2011

Development of a physically-based method for delineation of hydrologically homogeneous regions and flood quantile estimation in ungauged basins via the index flood method

Fredline Ilorme
Michigan Technological University

DEVELOPMENT OF A PHYSICALLY-BASED METHOD FOR DELINEATION OF HYDROLOGICALLY HOMOGENEOUS REGIONS AND FLOOD QUANTILE ESTIMATION IN UNGAUGED BASINS VIA THE INDEX FLOOD METHOD

By

Fredline Ilorme

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Civil Engineering

MICHIGAN TECHNOLOGICAL UNIVERSITY

2011

This dissertation, "Development of a Physically-Based Method for Delineation of Hydrologically Homogeneous Regions and Flood Quantile Estimation in Ungauged Basins via the Index Flood Method" is hereby approved in partial fulfillment of the requirements for the Degree of DOCTOR OF PHILOSOPHY IN CIVIL ENGINEERING.

Department of Civil and Environmental Engineering

Signatures:

Dissertation Advisor _____

Dr. Veronica W. Griffis

Department Chair _____

Dr. David Hand

Date _____

Table of Contents

Table of Contents 3
List of Figures 5
List of Tables 8
ACKNOWLEDGMENTS 13
ABSTRACT 14
Chapter 1 : Introduction 16
1.1 Research Motivation 18
1.2 Research Objectives and Organization of Chapters 21
Chapter 2 : Regional Flood Frequency Analysis 25
2.1 At-Site Flood Frequency Techniques 25
2.1.1 GEV Distribution 26
2.1.2 L-Moments 29
2.2 Regional Flood Frequency Techniques 31
2.2.1 Index Flood Method 32
2.2.1.1 Hydrological Homogeneity Test 34
2.2.1.2 Delineation of Homogeneous Regions 38
2.2.2 Generalized Least Squares Regression 48
Chapter 3 : A Standardized Procedure for Delineation of Hydrologically Homogeneous Regions 53
3.1 Standardized Procedure for Region Delineation 54
3.2 Data and Study Location 57
3.3 Attribute Selection 64
3.3.1 Statistics-Based Regions 64
3.3.2 Physically-Based Regions 68
3.3.3 Selected Attributes for Set 1 72
3.4 Region Delineation in Set 1 Using Standardized Procedure 72
3.5 Accuracy of Quantile Estimators for Ungauged Basins in Set 1 76
3.5.1 Results for Case 1 79
3.5.2 Results for Case 2 80

3.5.3 Results for Case 3 83
3.5.4 Results for Case 4 85
3.5.5 Results for Case 5 88
3.5.6 Comparison of Cases 89
3.6 Validation of Standardized Procedure in Alternate Study Areas....91
3.7 Conclusion 95
Chapter 4 : Delineation of Hydrologically Homogeneous Regions Using Spatially Distributed Data.... 97
4.1 Unsupervised Classification....99
4.2 Derivation of Spatially Distributed Variables....101
4.3 Region Delineation Using Spatially Distributed Attributes....109
4.4 Accuracy of Quantile Estimators for Ungauged Basins 114
4.5 Conclusion 118
Chapter 5 : Development of Quantile Estimators for Ungauged Basins in Data Sparse Areas ... 120
5.1 Estimation of the At-site Mean via GLS Regression....121
5.2 Flood Quantile Estimation at Ungauged Sites within the Predefined Study Area....126
5.3 Extension of GLS-IF Model to Alternate Areas 128
5.3.1 Validation via Extension to Data Rich Areas 129
5.3.2 Estimation at Ungauged Sites Located in Haiti 138
5.4 Conclusion 140
Chapter 6 : Conclusion.... 142
References.... 147
Appendix A.... 158
Appendix B.... 161
Appendix C.... 166
Appendix D.... 171
Appendix E 185

List of Figures

Figure 3.1: Location of unimpaired and unnested gauging stations within the Southeastern U.S. Sites in Set 1 comprise the training set; sites in Set 2 comprise the validation set. (State boundary files downloaded from Census Website: http://www.census.gov/geo/www/cob/st2000.html).....58

Figure 3.2: Dendrogram for Wards clustering applied to at-site L-CVs in Set 1. 66

Figure 3.3: Three Statistics-Based Regions delineated using Wards clustering applied to at-site L-CVs in Set 1. Circles represent HD sites, squares represent HPD sites. 66

Figure 3.4: Dendrogram for Wards clustering applied to E, S_B, and SI in Set 1. 74

Figure 3.5: Seven regions delineated for Set 1 using Wards clustering applied to normalized and standard values of E, S_B, and SI. Circles represent HD sites, and squares represent HPD sites. 75

Figure 3.6: MSE of flood quantiles obtained for regions delineated in Case 1. 80

Figure 3.7: Five regions delineated for Set 1 using Wards clustering applied to all nine normalized and standardized physical variables. Circles represent HD sites, and squares represent HPD sites. 81

Figure 3.8: MSE of flood quantiles obtained for regions delineated in Case 2. 82

Figure 3.9: Five regions delineated for Set 1 using Wards clustering applied to PC_1, PC_2 and PC_3 created using the nine normalized and standardized physical variables. Circles represent HD sites, and squares represent HPD sites. 84

Figure 3.10: MSE of flood quantiles obtained for regions delineated in Case 3. 85

Figure 3.11: Five regions delineated for Set 1 using Wards clustering applied to the first physical canonical variate. Circles represent HD sites and squares represent HPD sites. 87

Figure 3.12: MSE of flood quantiles obtained for regions delineated in Case 4. 88

Figure 3.13: MSE of flood quantiles obtained for regions delineated in Case 5. 89

Figure 3.14: Average MSE of flood quantiles obtained for regions delineated in Cases 1 through 5. Results are for regions appropriately modified to remove HPD sites. 91

Figure 3.15: Five regions delineated for Set 2 using Wards clustering applied to normalized and standardized values of E, S_B, and SI. Circles represent HD sites and squares represent HPD sites. 94

Figure 3.16: MSE of flood quantiles obtained for regions delineated in Set 2 as a function of E, SB, and SI. 95

Figure 4.1: Spatial representation of the fourteen soil drainage classes in the Southeastern U.S. as obtained from STATSGO (http://www.soilinfo.psu.edu/). 102

Figure 4.2: Spatial representation of eight slope classes for the Southeastern U.S. created using the ISODATA algorithm. 106

Figure 4.3: Spatial representation of fourteen elevation classes for the Southeastern U.S. created using the ISODATA algorithm. 107

Figure 4.4: Spatial extent of watersheds corresponding to gauging stations in Figure 3.1....109

Figure 4.5: Region delineation for Set 1 obtained using Wards clustering applied to spatially distributed values of elevation, basin slope, and soil drainage. Circles represent HD sites, and squares represent HPD sites. 111

Figure 4.6: Region delineation for Set 2 obtained using Wards clustering applied to spatially distributed values of basin elevation, basin slope, and soil drainage. Circles represent HD sites, and squares represent HPD sites. 113

Figure 4.7: MSE of flood quantiles obtained for regions delineated in Set 1 using spatially distributed representations of basin slope, elevation, and soil drainage. 116

Figure 4.8: MSE of flood quantiles obtained for regions delineated in Set 2 using spatially distributed representations of basin slope, elevation, and soil drainage. 116

Figure 4.9: Average MSE of flood quantiles obtained across regions delineated in Sets 1 and 2 using either aggregated or spatially distributed representations of basin slope, elevation, and soil drainage. 118

Figure 5.1: Correlation-distance smoothing function for Region 1 ($\alpha = 0.000053$, $\theta = 0.9998$). 123

Figure B.1: Distance plot obtained from the hydrological discordancy test of Neykov et al. (2007) and Station IDs of sites identified as hydrologically discordant. 165

Figure D.1: Correlation-distance smoothing function for Region 2 ($\alpha = 0.000083$ and $\theta = 0.9997$). 175

Figure D.2: Correlation-distance smoothing function for Region 3 ($\alpha = 0.000679$ and $\theta = 0.9994$) 175

Figure D.3: Correlation-distance smoothing function for Region 4 ($\alpha = 0.0003623$ and $\theta = 0.9996$). 176

Figure D.4: Correlation-distance smoothing function for Region 5 ($\alpha = 0.000245$ and $\theta = 0.9997$). 176

Figure D.5: Correlation-distance smoothing function for Region 6 ($\alpha = 0.000157$ and $\theta = 0.9998$). 177

Figure D.6: Correlation-distance smoothing function for Region 7 ($\alpha = 0.000565$ and $\theta = 0.9994$). 177

Figure D.7: Plot of residuals for GLS regression model of the mean for Region 1. 181

Figure D.8: Plot of residuals for GLS regression model of the mean for Region 2. 182

Figure D.9: Plot of residuals for GLS regression model of the mean for Region 3. 182

Figure D.10: Plot of residuals for GLS regression model of the mean for Region 4. 183

Figure D.11: Plot of residuals for GLS regression model of the mean for Region 5. 183

Figure D.12: Plot of residuals for GLS regression model of the mean for Region 6. 184

Figure D.13: Plot of residuals for GLS regression model of the mean for Region 7. 184

List of Tables

Table 3.1 Number of sites per state and record length statistics. 58

Table 3.2 Summary statistics for the physical variables in Sets 1 and 2 of the study area. 61

Table 3.3 Exponent for Box-Cox transformations of physical variables 62

Table 3.4 Summary statistics of the transformed (normalized) physical variables in Sets 1 and 2. 63

Table 3.5 Size and homogeneity of Statistics-Based Regions delineated using Wards clustering applied to at-site L-CVs in Set 1. 67

Table 3.6 Correlation coefficients resulting from LDA applied to normalized and standardized physical attributes of the Statistics-Based Regions in Set 1. 68

Table 3.7 Principal component loadings obtained for normalized and standardized physical variables at sites in Set 1 70

Table 3.8 Canonical correlations and coefficients of physical variates defined for Set 1 71

Table 3.9 Canonical correlations and coefficients of the hydrological variates defined for Set 1. 72

Table 3.10 Size and homogeneity of regions delineated for Set 1 using Wards clustering applied to normalized and standardized values of E, S_B, and SI 75

Table 3.11 Size and homogeneity of regions delineated for Set 1 using Wards clustering applied to all nine normalized and standardized physical variables. 82

Table 3.12 Size and homogeneity of regions delineated for Set 1 using Wards clustering applied to PC_1, PC_2 and PC_3 created using the nine normalized and standardized physical variables. 84

Table 3.13 Size and homogeneity of regions delineated for Set 1 using Wards clustering applied to the first physical canonical variate. 87

Table 3.14 Average heterogeneity (H*) obtained for regions delineated in Cases 1-5. 91

Table 3.15 Size and homogeneity of regions delineated for Set 2 using Wards clustering applied to normalized and standardized values of E, S_B, and SI.....94

Table 4.1 Characteristics of drainage classes for the Southeastern U.S. 104

Table 4.2 Characteristics of slope classes for Southeastern U.S. created using ISODATA algorithm. 107

Table 4.3 Characteristics of elevation classes for the Southeastern U.S. created using the ISODATA algorithm.....108

Table 4.4 Size and homogeneity of regions delineated using Wards clustering applied to spatially distributed values of basin elevation, basin slope, and soil drainage for sites in Set 1. 112

Table 4.5 Size and homogeneity of regions delineated using Wards clustering applied to spatially distributed values of basin elevation, basin slope, and soil drainage for sites in Set 2. 114

Table 5.1 OLS regression models of the coefficient of variation derived for each region in Set 1. 124

Table 5.2 GLS regression models of the at-site mean derived for each region in Set 1.....125

Table 5.3 Summary statistics for GLS regression models of the at-site mean derived for each region in Set 1. 126

Table 5.4 MSE of flood quantiles obtained using the GLS-IF approach for sites within Set 1. 128

Table 5.5 Coefficients of the normalized and standardized physical variables in the discriminant functions differentiating between regions in Set 1. 130

Table 5.6 Correlation coefficients resulting from LDA applied to the normalized and standardized physical variables for regions in Set 1. 130

Table 5.7 Number of sites (N) in Set 2 allocated to each region in Set 1 using LD_1. 132

Table 5.8 Confusion matrix for the classification scheme based on LD_1. 132

Table 5.9 MSE of quantiles estimated for sites in Set 2 using the Index Flood model for the corresponding region in Set 1 scaled by the at-site mean computed using the available record at the site of interest (Phase I).....135

Table 5.10 MSE of quantiles estimated for sites in Set 2 using the GLS-IF approach wherein the GLS estimator of the mean derived for the site of interest is used to scale the Index Flood model for the corresponding region in Set 1 (Phase II). 135

Table 5.11 Summary statistics for the mean annual precipitation (inches) for the seven regions delineated for Set 1 versus that of sites from Set 2 allocated to those regions. 137

Table 5.12 MSE of quantiles estimated for sites in Set 2 using GLS-IF(P) approach wherein the GLS estimator of the mean derived for the site of interest is scaled by the ratio of the at-site precipitation to the mean precipitation of the region in Set 1 within which it is allocated. 138

Table 5.13 Watershed Characteristics of Gonaives, Haiti. 139

Table 5.14 Flood quantiles derived for Gonaives watershed via the GLS-IF(P) approach. 140

Table B.1 PCA output obtained using normalized and standardized physical variables at sites in Region 2 of Set 1. 162

Table B.2 First four principal component scores for each watershed in Region 2 of Set 1. 163

Table B.3 Critical values for the first four principal component scores for Region 2 in Set 1. 163

Table C.1 MSE of flood quantiles obtained for regions delineated in Case 1 (Chapter 3). 166

Table C.2 MSE of flood quantiles obtained for regions delineated in Case 2 (Chapter 3). 167

Table C.3 MSE of flood quantiles obtained for regions delineated in Case 3 (Chapter 3). 167

Table C.4 MSE of flood quantiles obtained for regions delineated in Case 4 (Chapter 3). 168

Table C.5 MSE of flood quantiles obtained for regions delineated in Case 5 (Chapter 3). 168

Table C.6 MSE of flood quantiles obtained for regions delineated in Set 2 using aggregated values of basin elevation, basin slope, and soil drainage. 169

Table C.7 MSE of flood quantiles obtained for regions delineated in Set 1 using spatially distributed representations of basin elevation, basin slope, and soil drainage. 169

Table C.8 MSE of flood quantiles obtained for regions delineated in Set 2 using spatially distributed representations of basin elevation, basin slope, and soil drainage. 170

Table D.1 Summary statistics for OLS regression model coefficients for each region in Set 1. 173

Table D.2 Summary statistics for the OLS regression models of the coefficient of variation derived for each region in Set 1. 174

Table D.3 Summary statistics for GLS regression model coefficients in Region 1. 178

Table D.4 Summary statistics for GLS regression model coefficients in Region2. 179

Table D.5 Summary statistics for GLS regression model coefficients in Region 3. 179

Table D.6 Summary statistics for GLS regression model coefficients in Region 4. 179

Table D.7 Summary statistics for GLS regression model coefficients in Region 5. 180

Table D.8 Summary statistics for GLS regression model coefficients in Region 6. 180

Table D.9 Summary statistics for GLS regression model coefficients in Region 7. 181

Table E.1 Summary statistics for the nine physical variables computed for sites within Region 1 of Set 1 versus that of sites from Set 2 allocated therein. ...186

Table E.2 Summary statistics for the nine physical variables computed for sites within Region 2 of Set 1 versus that of sites from Set 2 allocated therein. ...187

Table E.3 Summary statistics for the nine physical variables computed for sites within Region 3 of Set 1 versus that of sites from Set 2 allocated therein. ...187

Table E.4 Summary statistics for the nine physical variables computed for sites within Region 4 of Set 1 versus that of sites from Set 2 allocated therein. ...188

Table E.5 Summary statistics for the nine physical variables computed for sites within Region 5 of Set 1 versus that of sites from Set 2 allocated therein. ...188

Table E.6 Summary statistics for the nine physical variables computed for sites within Region 6 of Set 1 versus that of sites from Set 2 allocated therein. ...189

Table E.7 Summary statistics for the nine physical variables computed for sites within Region 7 of Set 1 versus that of sites from Set 2 allocated therein. ...189

Table E.8 p-values of Wilcoxon-Mann-Whitney test. 190

Table E.9 Confusion matrix for the classification scheme based on the first two discriminant functions. 191

Table E.10 Confusion matrix for the classification scheme based on the first three discriminant functions. 191

Table E.11 Confusion matrix for the classification scheme based on the first four discriminant functions. 192

Table E.12 Confusion matrix for the classification scheme based on the first five discriminant functions. 192

Table E.13 Confusion matrix for the classification scheme based on all six discriminant functions. 193

ACKNOWLEDGMENTS

I wish to thank my advisor, Dr. Veronica Griffis, for her guidance and advice throughout the course of this dissertation. Her input has been invaluable and contributed immensely to the success of this research. I also wish to thank my committee members: Dr. David Watkins, Dr. John Gierke, Dr. Remigio Galárraga-Sánchez, and Dr. Huann-Sheng Chen for their patience and astute comments. My gratitude also extends to the Sustainable Futures Institute, the Civil and Environmental Engineering department and the Michigan Technological University Graduate School for their generous support. Special thanks to my friends: Essa, Petra and Miriam as well as my husband and my family members in Haiti and in the States. This dissertation would not be possible without your faith and encouragements. Thank You.

ABSTRACT

Regional flood frequency techniques are commonly used to estimate flood quantiles when flood data is unavailable or the record length at an individual gauging station is insufficient for reliable analyses. These methods compensate for limited or unavailable data by pooling data from nearby gauged sites. This requires the delineation of hydrologically homogeneous regions in which the flood regime is sufficiently similar to allow the spatial transfer of information. It is generally accepted that hydrologic similarity results from similar physiographic characteristics, and thus these characteristics can be used to delineate regions and classify ungauged sites. However, as currently practiced, the delineation is highly subjective and dependent on the similarity measures and classification techniques employed.

A standardized procedure for delineation of hydrologically homogeneous regions is presented herein. Key aspects are a new statistical metric to identify physically discordant sites, and the identification of an appropriate set of physically based measures of extreme hydrological similarity. A combination of multivariate statistical techniques applied to multiple flood statistics and basin characteristics for gauging stations in the Southeastern U.S. revealed that basin slope, elevation, and soil drainage largely determine the extreme hydrological behavior of a watershed. Use of these characteristics as similarity measures in the standardized approach for region delineation yields regions which are more homogeneous and more efficient for quantile estimation at ungauged sites than those delineated using alternative physically-based procedures typically

employed in practice. The proposed methods and key physical characteristics are also shown to be efficient for region delineation and quantile development in alternative areas composed of watersheds with statistically different physical composition. In addition, the use of aggregated values of key watershed characteristics was found to be sufficient for the regionalization of flood data; the added time and computational effort required to derive spatially distributed watershed variables does not increase the accuracy of quantile estimators for ungauged sites.

This dissertation also presents a methodology by which flood quantile estimates in Haiti can be derived using relationships developed for data rich regions of the U.S. As currently practiced, regional flood frequency techniques can only be applied within the predefined area used for model development. However, results presented herein demonstrate that the regional flood distribution can successfully be extrapolated to areas of similar physical composition located beyond the extent of that used for model development provided differences in precipitation are accounted for and the site in question can be appropriately classified within a delineated region.

Chapter 1 : Introduction

Accurate flood quantile estimates are necessary for the delineation of floodplains, the development of floodplain management and flood warning systems, and the design and operation of water-control structures, such as reservoirs and culverts. Standard procedures for at-site flood frequency analysis involve assembling the annual maximum flood record at the site of interest and fitting an analytic probability distribution to the data (e.g., IACWD 1982). The fitted distribution is then used to estimate flood quantiles associated with a given return period, such as the flood magnitude expected to be equaled or exceeded once every 100 years (i.e., the 100-year event). However, in most cases the at-site record length is too short to accurately estimate flood quantiles for return periods of interest: estimation of the 100-year event often requires extrapolating beyond the observed flood record. In other cases, flood data are unavailable at the site of interest, making at-site flood frequency analysis impossible. As the latter is often the case for watersheds throughout the world, particularly in data sparse developing countries, but also in data rich countries such as the United States (e.g., Mishra and Coulibaly 2009), the development of appropriate methods for flood quantile estimation in ungauged basins is a common research theme in hydrology.

To compensate for limited or unavailable flood data, one solution is to "trade space for time" (Stedinger et al. 1993) using a regional flood frequency analysis, wherein the characterization of flood flows at the site of interest is derived using information pooled from nearby hydrologically similar gauged sites (NRC 1988). Regional flood frequency methods include the Index Flood method (e.g. Dalrymple 1960; Hosking and

Wallis 1988, 1997; Stedinger and Lu 1995; Fill and Stedinger 1998; De Michele and Rosso 2001; Kjeldsen and Rosberg 2002), and regional regression procedures, such as weighted and generalized least squares regression (e.g. Tasker and Stedinger 1989; Tasker et al. 1996; Madsen and Rosberg 1997; Eng et al. 2005, 2007a, 2007b; Griffis and Stedinger 2007a; Jeong et al. 2007).

Much of the recent research has focused on improving or comparing existing regional flood frequency techniques (e.g., Castellarin et al. 2001; Chiang et al. 2002a; Kjeldsen and Rosbjerg 2002; Eng et al. 2007b; Griffis and Stedinger 2007a; Neykov et al. 2007; Gruber and Stedinger 2008), and developing new methods for quantile estimation at ungauged basins located within the area used for model development (e.g., Chiang et al. 2002b; Eng et al. 2005, 2007b; Kayha et al. 2008; Shu and Ouarda 2008; Saf 2009; Malekinezhad et al. 2010). The research presented in this dissertation draws on this base of knowledge to propose additional recommendations to improve quantile estimators for use within data rich areas, and proposes a novel method to extrapolate those results to sites with similar physical characteristics located in data sparse regions external to the area used for model development. In particular, this research presents a method by which the flood regime in Haiti can be derived based on knowledge of the relationships between flood statistics and physical characteristics within the Southeastern portion of the United States.

1.1 Research Motivation

An important component of the Index Flood method is the delineation of hydrologically homogeneous regions (or groups of sites), in which the flood regime is deemed sufficiently similar to allow the spatial transfer of information from gauged sites to ungauged sites (see for example, Stedinger and Lu, 1995, and citations therein). The flood regime includes the magnitude, timing, duration, frequency, and inter-annual variability of flood events. Within a hydrologically homogeneous region, sites share the same parent flood distribution with a common shape parameter, but each watershed has a site-specific scale factor (or location parameter). For application at ungauged sites, it is assumed that the flood regime, or extreme hydrologic response, is similar in watersheds with comparable basin characteristics (e.g. physiographic characteristics and meteorological inputs). Thus, a region in which sites are physically homogeneous is assumed to also be hydrologically homogenous, and physical characteristics can be used to classify an ungauged site within a delineated region. Some studies have observed, however, that similar hydrologic response is not guaranteed simply by similar basin characteristics because of complex interactions among those characteristics (Zrinji and Burn 1994; Burn 1997; Burn et al. 1997). In addition, delineated regions are highly dependent on the choice of similarity measures used to infer homogeneity (see for example, Burn 1990, 1997; GREHYS 1996; Burn et al. 1997; Hosking and Wallis 1997; Castellarin et al. 2001). These are major limitations of the Index Flood method, and significantly impair its ability to provide accurate quantile estimates at ungauged sites.

Regional regression models are also hindered by our limited knowledge of the physical properties and mechanisms producing flood flows. The typical modeling approach is to employ a log-log or log-linear relationship between flood statistics and basin characteristics, and in most cases the drainage area is the only explanatory variable employed. More complex relationships and interactions between basin characteristics are generally not considered because the current understanding of hydrology is not advanced enough to know if such a model is reasonable (Eng et al. 2007b; Griffis and Stedinger 2007a). Thus, models of flood quantiles and other statistics at ungauged sites can be improved by increasing our knowledge of what basin characteristics, or interactions among basin characteristics, predominantly contribute to flood flows.

With increasing computing abilities and the availability of remotely sensed data, it may also be possible to improve quantile estimates in data limited areas by using available data more efficiently. Recent research demonstrates the ability to delineate climatic regions as a function of remotely sensed data, including land surface temperature, precipitation, and infiltration categories based on microtopography, surface crusting and soil cover (Corbane et al. 2008; Rhee et al. 2008). Remote sensing systems have also been used extensively to identify soil type, land use, land cover, geology and topography (Corbane et al. 2008; Brink and Eva 2009; Bertoldi et al. 2010; Inbar et al. 2010). Previous studies suggest that geology, land use, and land cover may help define the flood distribution of drainage basins (Chiang et al. 2002a, 2002b; Rao et al. 2006), and thus remotely sensed data could be used to infer the flood regime in areas with limited or unavailable flood data.

Regionalization of flood data using either the Index Flood method or regional regression assumes that the watershed processes governing the flood regime are sufficiently characterized by physical parameters aggregated at the watershed scale. Some may argue that spatially distributed parameters should be used to develop finer scale representations of hydrological processes (Beven and Kirkby 1979; Abbott et al. 1986; Boyle et al. 2001; Duffy 2004; Panday and Huyakorn 2004; Reed et al. 2007). However, there is a trade-off between characterizing the heterogeneity within and uniqueness of a single watershed using spatially distributed values as commonly practiced in hydrologic modeling, and characterizing the heterogeneity within a region using parameter values aggregated at the watershed scale as in regional flood frequency techniques. In the latter analyses, simple models are needed to infer the dominant processes governing extreme hydrologic response at the watershed scale, such that flood statistics can be successfully extrapolated from gauged basins for improved prediction in ungauged basins in data limited areas. (See for example, McDonnell et al. 2007; Tezlaff et al. 2008; MacKinnon and Tetzlaff 2009, and citations therein.) Use of these simple models in conjunction with remotely sensed data would allow for the development of quantile estimators in data sparse countries such as Haiti by extrapolating the relationships developed for basins of similar physical composition in data rich countries such as the United States.

Hydrologic models such as rainfall-runoff models are another option for flood quantile estimation in ungauged basins. Unlike regional flood frequency analyses, however, rainfall-runoff models do not provide information pertaining to the flood

distribution nor do they explain possible similarities in flood distributions among watersheds. Further, rainfall-runoff models pose additional problems because the critical storm duration and the spatial distribution of relevant storm events (and corresponding inflows) are unknown. In general, regional flood frequency analyses provide less accurate flood quantile estimates than at-site flood frequency analyses when sufficient gauged data is available (see for example, Griffis and Stedinger 2007b), but often provide more accurate flood quantiles than hydrologic modeling (USACE 1994).

1.2 Research Objectives and Organization of Chapters

The objectives of the research presented herein are: (i) to increase understanding of the basin characteristics, or interactions among basin characteristics, which predominantly contribute to flood flows, (ii) to improve estimates of flood quantiles at ungauged sites by removing subjectivity from the process of delineating hydrologically homogeneous regions, and (iii) to develop simple rules to successfully extrapolate flood statistic-basin characteristic relationships from data rich areas to derive flood quantile estimates in data sparse areas outside of that used for region delineation and model development. In particular, this dissertation seeks to develop methods by which the flood regime in Haiti can be derived based on knowledge of the relationships between flood statistics and physical characteristics within the Southeastern U.S. These objectives are achieved via completion of four tasks as outlined below.

Task 1: Evaluate to what extent hydrological homogeneity is explained by physical homogeneity, and identify the key physical characteristics needed to delineate

hydrologically homogeneous regions. This is accomplished using a combination of traditional regional flood frequency techniques and multivariate statistical analyses applied to multiple flood statistics and watershed characteristics for sites throughout the Southeastern U.S. Chapter 2 provides an overview of relevant flood frequency techniques and multivariate statistics. Chapter 3 discusses the application of these techniques to two study areas within the Southeastern U.S. The first study area is used to identify the physical attributes which are most indicative of the flood regime, and the second study area is used to validate the proposed application of the identified physical attributes for region delineation and inference of the flood regime in areas outside of the extent of the first study area.

Task 2: Develop a standardized procedure for the delineation of hydrologically homogeneous regions. The proposed method, including a new statistical procedure to identify physically discordant sites, is presented in Chapter 3. Data for sites across the Southeastern U.S. are used to demonstrate the application of the proposed method, and to evaluate the accuracy of quantile estimators for ungauged sites derived using the proposed method relative to estimators derived using procedures typically employed in practice.

***Task 3*:** Demonstrate that values of physical variables aggregated at the watershed scale are sufficient for the delineation of regions suitable for flood frequency analysis. This is accomplished in Chapter 4 wherein hydrologically homogeneous regions are delineated within the Southeastern U.S. using a novel approach based on spatially distributed representations of the key physical characteristics identified in Task 1.

Quantile estimators for ungauged sites are subsequently derived and compared to estimators derived in Chapter 3 based on aggregated parameter values. Both sets of quantile estimators are derived for regions delineated using the standardized method developed in Task 2; the procedures differ only in their use of spatially distributed values versus aggregated values of the key physical characteristics.

***Task 4*:** Demonstrate that quantile estimators derived for data rich regions can be extrapolated to data sparse areas which have similar physical characteristics, but are located outside of the area used for model development. This is accomplished in Chapter 5 using a combination of GLS regression models and the Index Flood method applied in regions delineated using the standardized procedure developed in Task 2. The success of the extrapolation is first evaluated using data for sites across the Southeastern U.S. The dataset is split into two study areas: one represents the data rich area used for model development; the second represents the data poor region. Applications which demonstrate the extrapolation to data sparse countries are then performed for sites in Haiti.

Overall, this research provides a critical contribution to flood frequency analysis as appropriate methods for flood quantile estimation in ungauged basins are needed for data sparse regions of the world. For instance, the necessary data for flood frequency analysis is often unavailable in developing countries due to lack of financial resources or for political reasons, however, remotely sensed data is now freely and readily available for most of the world. In such cases, remote sensing systems provide a way to obtain data where little information was previously available. In addition, advances in

computational capabilities make remote sensing data coupled with traditional regional flood frequency techniques quite attractive and less time consuming than alternatives such as the development of watershed specific rainfall-runoff models. Although synthetic unit hydrographs can simplify the process, their applicability to the watershed of interest may still be questionable.

Chapter 2 : Regional Flood Frequency Analysis

Flood series at an individual site are seldom long enough to accurately estimate flood quantiles for return periods of interest. In other cases, flood data are unavailable at the site of interest, making at-site flood frequency analysis impossible. Regional flood frequency techniques which employ data from nearby sites have thus been developed to overcome the lack of flood data at a particular location. This chapter reviews standard at-site and regional flood frequency techniques and discusses limitations of their application to ungauged basins.

2.1 At-Site Flood Frequency Techniques

Standard procedures for estimating the risk of flood events involve estimating the return period associated with the magnitude of observed events using an analytic probability distribution. Once properly fit to the observed data, the distribution is then used to determine the magnitude of needed design events (e.g., the 100-year event) for use in water resources applications such as land-use planning and management, and the design and operation of water-use and water-control structures. As the true distribution of annual maximum flood flows is unknown, this process requires the selection of an analytic probability distribution that reasonably approximates the observations and their corresponding recurrence intervals, as well as an appropriate parameter estimation technique to fit the distribution using the available period of record.

Distributions commonly used to model annual maximum flood series are the log-Pearson type III (LP3), generalized extreme value (GEV), and lognormal (LN)

distributions (Stedinger et al. 1993). Vogel and Wilson (1996) demonstrate that all three of these distributions are reasonable models of annual maximum flood series in the United States; however, the LP3 distribution is typically used to model flood series as recommended by *Bulletin 17B* (IACWD 1982). Further, Griffis and Stedinger (2007c) demonstrate that, with two shape parameters, the LP3 distribution is more flexible than the LN and GEV distributions. However, when compared to the uncertainty in the flood quantile estimates, the differences between reasonable choices of distributions are negligible (Stedinger 1980; Hosking and Wallis 1997; Stedinger and Griffis 2008). Therefore, the GEV distribution will be employed herein, because it has been widely used throughout the world (e.g., NERC 1975) and has recently gained a lot of support for at-site flood frequency in the U.S. (e.g., Saf 2010; Malekinezhad et al. 2011), and it is frequently used in the context of regional index flood modeling (e.g., Hosking et al. 1985b; Wallis and Wood 1985; Lettenmaier et al. 1987; Hosking and Wallis 1988; Chowdhury et al. 1991; Stedinger and Lu 1995; Hosking and Wallis 1997; Madsen et al. 1997). The GEV distribution and relevant parameter estimation methods are discussed in more detail below.

2.1.1 GEV Distribution

Extreme value theory demonstrates that regardless of the distribution of Y, if n is large enough, then the distribution of $X_n = \max \{Y_1, \ldots, Y_n\}$ converges to either the Gumbel, Frechet, or Weibull distribution (Gumbel 1958). These distributions differ primarily in regard to their behavior in the tails. The GEV distribution combines these three distributions into one distribution (Jenkinson 1969). The GEV distribution can be

described by the location parameter ξ, scale parameter α and shape parameter k , and the cumulative distribution function (cdf):

$$F(X) = exp\left\{-\left[1 - \frac{k(x-\xi)}{\alpha}\right]^{\frac{1}{k}}\right\} \quad (1)$$

The shape parameter k defines the tail behavior of the distribution. When $k < 0$, the GEV distribution corresponds to the Frechet distribution with a heavy right-hand tail and a lower bound equal to $\xi + \alpha/k$; when $k > 0$, the GEV distribution corresponds to the Weibull distribution with an upper bound equal to $\xi + \alpha/k$; and when $k = 0$, the GEV distribution reduces to the Gumbel distribution with an exponential upper tail and the following cdf:

$$F(X) = exp\left\{-exp\left[-\frac{x-\xi}{\alpha}\right]\right\} \quad (2)$$

which is unbounded both above and below ($-\infty < x < \infty$).

For k > -1/3, the GEV parameters (ξ, α, and k) are functions of the first three population moments (mean μ_x, variance σ_x^2, and skew γ_x):

$$\mu_x = \xi + \frac{\alpha}{k}[1 - \Gamma(1+k)]$$

$$\sigma_x^2 = \left(\frac{\alpha}{k}\right)^2 \{\Gamma(1+2k) - [\Gamma(1+k)]^2\}$$

$$\gamma_x = sign(k)\frac{-\Gamma(1+3k) + 3\Gamma(1+k)\Gamma(1+2k) - 2\left[\Gamma(1+k)\right]^3}{\left\{\Gamma(1+2k) - \left[\Gamma(1+k)\right]^2\right\}^{3/2}} \quad (3)$$

where sign(k) = +/- 1 depending on the sign of k, and $\Gamma(c)$ is the gamma function

$$\Gamma(c) = \int_0^\infty t^{c-1} \exp(-t)dt \tag{4}$$

Quantiles of the GEV distribution are computed using the following equation:

$$\begin{aligned} x_p &= \xi + \frac{\alpha}{k}\{1 - [-\ln(p)]^k\} \qquad \text{for} \quad k \neq 0 \\ &= \xi - \alpha \ln[-\ln(p)]^k \qquad \text{for} \quad k = 0 \end{aligned} \tag{5}$$

where p is the cumulative probability corresponding to the return period of interest, T = 1/(1-p). The p^{th} quantile (x_p) is thus the flood magnitude expected to be equaled or exceeded once every T years. In practice, quantile estimates ($\hat{x}_p$) are obtained by replacing the true parameter values (ξ, α, k) with parameter estimators ($\hat{\xi}$, $\hat{\alpha}$, $\hat{k}$) computed as a function of the observed flood series.

An efficient method for parameter estimation is necessary when using the GEV distribution in flood frequency applications as the record length at an individual site is often limited. The most common methods used to estimate the parameters of the GEV distribution are: (i) method of moments (MOM) which equates the population moments in equation (3) to the sample moments ($\overline{x}$, s_x^2, and G_x) computed using traditional moment estimators; (ii) maximum likelihood estimation (MLE) which identifies the values of the parameters that maximize the likelihood of having observing the flood series in question; and, (iii) L-moments estimators which are based on linear combinations of the ranked data. Martins and Stedinger (2000) provide a detailed discussion of each of these methods with respect to application of the GEV distribution in flood frequency analysis. Overall, L-moments estimators have been shown to be the

most efficient parameter estimation method for the GEV distribution (e.g., Landwehr et al. 1979; Hosking et al. 1985a) and will be used in subsequent analyses.

2.1.2 L-Moments

L-moments are an alternative to traditional product moments (e.g., $\overline{x}$, s_x^2, and G_x) for describing the statistical properties of flood data. L-moments are computed as functions of probability weighted moments (PWMs) defined as

$$\beta_r = E\{X\,[F(X)]^r\} \tag{6}$$

where *F(X)* is the cumulative distribution function for *X* (Greenwood et al. 1979; Hosking, 1990). L-moments are then calculated using the following relationships

$$\lambda_1 = \beta_0 \tag{7}$$

$$\lambda_2 = 2\beta_1 - \beta_0 \tag{8}$$

$$\lambda_3 = 6\beta_2 - 6\beta_1 + \beta_0 \tag{9}$$

$$\lambda_4 = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0 \tag{10}$$

The first L-moment (λ_1) is equivalent to the population mean (μ). A dimensionless L-moment coefficient of variation (L-CV) is given by

$$\tau_2 = \frac{\lambda_2}{\lambda_1} \tag{11}$$

The L-moment coefficient of skewness (L-Skewness) and kurtosis (L-Kurtosis) are given by

$$\tau_r = \frac{\lambda_r}{\lambda_2} \tag{12}$$

for r = 3 and 4, respectively. These dimensionless L-moment coefficients, commonly referred to as L-moment ratios, are analogues for the conventional coefficients of variation (CV = $s_x / \overline{x}$), skewness and kurtosis.

In practice, sample L-moments ($\hat{\lambda}_i$) and L-moment ratios ($\hat{\tau}_r$) are obtained by replacing the population PWMs in equations (7) - (12) with the sample PWMs computed using

$$\hat{\beta}_r = b_r = \frac{1}{n}\sum_{j=1}^{n-r} \frac{\binom{n-j}{r} X_{(j)}}{\binom{n-1}{r}} \tag{13}$$

(Landwehr et al., 1979). This formula yields an unbiased estimator of the order r PWM for a sample of size n ranked in descending order ($X_{(n)} \leq \cdots .. \leq X_{(1)}$). In this way, L-moments are based on linear combinations of the observations, as opposed to product moments which involve squaring and cubing the observations, (Hosking, 1990; Hosking and Wallis, 1997). As such, parameter estimators obtained using the product-moment sample coefficients of variation, skewness, and kurtosis are highly variable and have a bias which depends upon the sample size as well as the underlying distribution (Wallis 1988; Wallis et al. 1974). In fact, for a sample of size n, the product-moment sample coefficient of variation cannot be larger than $(n-1)^{0.5}$, and the sample skewness cannot exceed $(n-2)/(n-1)^{0.5}$ (Kirby 1974). Conversely, L-moments computed using unbiased PWMs yield unbiased analogues for the conventional coefficients of variation, skewness and kurtosis, and the resulting parameter estimators are more efficient in smaller samples than those derived from traditional product moments (Wallis 1988).

In applications of the GEV distribution, the parameters can be computed as a function of the L-moments using

$$k = 7.8590c + 2.9554c^2 \tag{14}$$

$$\alpha = \frac{k\lambda_2}{\Gamma(1+k)\left(1-2^{-k}\right)} \tag{15}$$

$$\xi = \lambda_1 + \frac{\alpha}{k[\Gamma(1+k)-1]} \tag{16}$$

where the expression for k is an approximation developed by Hosking et al. (1985[a]) with

$$c = \frac{2\lambda_2}{\lambda_3+3\lambda_2} - \frac{\ln(2)}{\ln(3)} \tag{17}$$

Aside from parameter estimation in at-site analyses, the L-moment ratios are also used in regional flood frequency analyses to ascertain the homogeneity of a collection of watersheds and to derive quantile estimates at sites with limited data as discussed below.

2.2 Regional Flood Frequency Techniques

The two most common regional flood frequency techniques are the Index Flood method (e.g. Dalrymple 1960; Hosking and Wallis 1988, 1997; Stedinger and Lu 1995; Fill and Stedinger 1998; De Michele and Rosso 2001; Kjeldsen and Rosberg 2002), and regional regression analyses (e.g. Tasker and Stedinger 1989; Tasker et al. 1996; Madsen and Rosberg 1997; Eng et al. 2005, 2007a, 2007b; Griffis and Stedinger 2007a; Jeong et al. 2007). Each of these procedures and their application in the context of the research herein are discussed below.

2.2.1 Index Flood Method

The Index Flood method is based on the premise that sites within a statistically (or hydrologically) homogeneous region share the same parent (or regional) flood frequency distribution with a common shape parameter, but each watershed has a site-specific scale parameter (a.k.a. the "index-flood") to represent possible changes in magnitude across the region. For application at gauged sites, the scale parameter is often given by the mean of the flood flows. For ungauged sites, however, this parameter must be related to physiographic characteristics of the watershed, the most important of which is drainage area. This could be accomplished using regional regression procedures as discussed in Section 2.2.2. The basin characteristics at any point in the region (i.e. an ungauged site) can then be used to estimate the mean annual flood, which in turn can be used with the non-dimensional parent distribution to determine the flood magnitude corresponding to any return period at that location.

While any combination of probability distribution and parameter estimation method could be used, the use of the GEV distribution fit using L-moments has been shown to produce more accurate quantile estimates than other distribution/estimation method combinations (e.g., Hosking and Wallis 1988; Jin and Stedinger 1989; Potter and Lettenmaier 1990; Rosbjerg and Madsen 1995; Stedinger and Lu 1995). The GEV/L-moment Index Flood procedure for application at a gauged site within a predefined hydrologically homogeneous region is as follows (Stedinger et al. 1993):

1) Calculate the L-moment estimators $\hat{\lambda}_1^{(j)}, \hat{\lambda}_2^{(j)}$, and $\hat{\lambda}_3^{(j)}$ at site j using the unbiased PWM estimators b_0, b_1, and b_2 in equations (7), (8), and (9). Repeat for all sites in the region (j = 1, 2…, N).

2) Define the non-dimensional parent distribution by the normalized regional L-moments ($\hat{\lambda}_1^R$, $\hat{\lambda}_2^R$, and $\hat{\lambda}_3^R$) computed using:

$$\hat{\lambda}_r^R = \frac{\sum_{j=1}^{N} w_j [\hat{\lambda}_r^{(j)} / \hat{\lambda}_1^{(j)}]}{\sum_{j=1}^{N} w_j} \qquad \text{for } r = 2, 3 \tag{18}$$

where $w_j = \frac{n_j n_R}{n_j + n_R}$, n_j is the record length at site *j*, and $n_R \approx \frac{\sum n_j}{N}$. For r = 1, $\hat{\lambda}_1^R = 1$.

3) Estimate the parameters ($\hat{\xi}^R$, $\hat{\alpha}^R$, and $\hat{k}^R$) and quantiles ($\hat{x}_p^R$) of the non-dimensional regional GEV distribution by using the normalized regional L-moments ($\hat{\lambda}_1^R$, $\hat{\lambda}_2^R$, and $\hat{\lambda}_3^R$) in equations (14) - (17) and (5), respectively.

4) Estimate the 100*p* percentile of the flood distribution, corresponding to the T = 1/(1-p) year event, at any site *j* using:

$$\hat{x}_p(j) = \hat{\lambda}_1^j \hat{x}_p^R \tag{19}$$

where $\hat{\lambda}_1^j$ is the at-site sample mean for site *j* computed using equation (7).

This procedure is meant to be applied in hydrologically homogeneous regions which consist of a group of sites for which the extreme hydrologic response within the corresponding watersheds is deemed sufficiently similar to allow the spatial transfer of information (e.g., Stedinger and Lu 1995, and citations therein). However, even when moderate regional heterogeneity is present, a regional analysis will still yield much more accurate extreme quantile estimates than an at-site analysis where limited gauged data is available (Lettenmaier et al. 1987; Hosking and Wallis 1988; Hosking and Wallis 1997). Methods typically applied in practice to group similar sites and subsequently evaluate their regional homogeneity are discussed in the following sections.

2.2.1.1 Hydrological Homogeneity Test

The hydrological homogeneity of a region requires that the sites within are able to share a common parent distribution, and thus the at-site distributions are similar in shape. Statistical similarity among the at-site distributions can therefore be assessed in terms of sample product moments such as the coefficient of variation (e.g., Dalrymple 1960), normalized flood quantile estimates (e.g., Lu and Stedinger 1992), or L-moment ratios such as the L-CV and L-Skewness (e.g., Hosking and Wallis 1997). The latter test on L-Moment ratios is most commonly used in practice, and will be employed herein as it is consistent with the use of the GEV/L-Moment Index Flood procedure.

The Hosking and Wallis (1997) homogeneity test compares the variability in the observed at-site L-moment ratios to the expected variability in the L-moment ratios based on simulations from representative hydrologically homogeneous regions consistent with

the observed regional averages of the L-moment ratios and available record lengths. Three possible measures of variability in the at-site L-moment ratios are:

$$V_1 = \left\{ \sum_{i=1}^{N} n_i \left(\hat{\tau}_2^{(i)} - \bar{\tau}_2 \right)^2 / \sum_{i=1}^{N} n_i \right\}^{1/2} \tag{20}$$

$$V_2 = \left\{ \sum_{i=1}^{N} n_i \left[\left(\hat{\tau}_2^{(i)} - \bar{\tau}_2 \right)^2 + \left(\hat{\tau}_3^{(i)} - \bar{\tau}_3 \right)^2 \right]^{1/2} / \sum_{i=1}^{N} n_i \right\}^{1/2} \tag{21}$$

$$V_3 = \left\{ \sum_{i=1}^{N} n_i \left[\left(\hat{\tau}_3^{(i)} - \bar{\tau}_3 \right)^2 + \left(\hat{\tau}_4^{(i)} - \bar{\tau}_4 \right)^2 \right]^{1/2} / \sum_{i=1}^{N} n_i \right\}^{1/2} \tag{22}$$

where $\hat{\tau}_2^{(i)}$, $\hat{\tau}_3^{(i)}$, and $\hat{\tau}_4^{(i)}$ are estimates of the L-CV, L-Skewness and L-Kurtosis at site i, $\bar{\tau}_2$, $\bar{\tau}_3$, and $\bar{\tau}_4$ are the regional averages of the L-moment ratios, n_i is the record length at site i, and N is the number of sites in the region. It has been observed that the first homogeneity measure (V_1) based solely on the L-CV is the most effective at discriminating between homogeneous and heterogeneous regions (Hosking and Wallis 1997; Castellarin et al. 2001), and therefore will be used herein as an indicator of hydrological homogeneity.

To compare the observed variability in L-CVs to what would be expected in a homogeneous region, Hosking and Wallis (1997) developed the H-statistic defined as:

$$H = \frac{(V_1 - \mu_V)}{\sigma_V} \tag{23}$$

Here V_1 represents the observed variability in the at-site L-CVs computed using equation (20). The expected variability should the region of interest be homogenous is represented by the mean μ_V and standard deviation σ_V of values of V_1 computed for a large number of

simulated regions consisting of uncorrelated sites with record lengths and regional average L-moment ratios equivalent to those of the observed data set. These simulations can be easily performed in R using the HOMTEST package (Viglione 2010). If $H < 1$, the region is considered hydrologically homogeneous, meaning the observed variability in the L-CVs is no larger than would be expected given natural sampling variability. If $1 \leq H < 2$, the region is possibly heterogeneous, but is generally considered sufficient for application of the Index Flood procedure. If $H \geq 2$, the region is definitely heterogeneous and steps should be taken to improve the homogeneity of the region, if possible. Results of recent studies suggest relaxing the latter constraint to $H \geq 4$ (Hosking and Wallis 1993; Robson and Reed 1999; Guse et al. 2009) given that the significance levels are only accurate in the absence of serial and cross correlation, and that the sample truly follows a kappa distribution (Viglione et al. 2007). In the study presented herein, all regions with $H < 4$ are considered appropriate for regionalization of flood data.

For regions deemed definitely heterogeneous ($H \geq 4$ herein), Hosking and Wallis (1997) recommend identifying possible discordant sites which may be responsible for inflating the H-statistic. Discordant sites are statistical outliers with L-moment ratios which are substantially different from the regional average L-moment ratios. These outliers can be identified using a discordancy measure computed for a given site as follows

$$D_i = \frac{1}{3} N (u_i - \bar{u}) A^{-1} (u_i - \bar{u})^T \tag{24}$$

where u_i is a vector of the at-site L-moment ratios $\left(\hat{\tau}_2^{(i)}, \hat{\tau}_3^{(i)}, \hat{\tau}_4^{(i)}\right)$ at site i, $\bar{u}$ is a vector of the regional average L-moment ratios ($\bar{\tau}_2$, $\bar{\tau}_3$, and $\bar{\tau}_4$), and

$$A = \sum_{i=1}^{N}(u_i - \bar{u})\,(u_i - \bar{u})^T \tag{25}$$

For a region containing at least 15 sites, site i is hydrologically discordant if D_i is greater than 3. Hydrologically discordant sites should be removed from the region only if they are also discordant on physical and/or climatological grounds (Hosking and Wallis 1997, p. 170). When necessary, regions may also be broken into smaller subsets.

To reduce the impact of outliers on the discordancy measure, Neykov et al. (2007) propose a robust discordancy measure computed using a subset of size h selected from the N sites available in the region. The selected subset has the minimum covariance determinant relative to all other possible subsets of size h. For a given site i in the subset, the robust discordancy measure is computed as

$$RD_i{}^2 = (u_i - L)C^{-1}(u_i - L)^T \tag{26}$$

where u_i is a vector of the at-site L-moment ratios $\left(\hat{\tau}_2^{(i)}, \hat{\tau}_3^{(i)}, \hat{\tau}_4^{(i)}\right)$ at site i, L is a vector of the means of the L-moment ratios ($\bar{\tau}_2$, $\bar{\tau}_3$, and $\bar{\tau}_4$) computed for sites within the subset, and

$$C = c_m \frac{1}{h}(u_i - L)(u_i - L)^T \tag{27}$$

The factor c_m is chosen to be consistent with the multivariate normal model and is unbiased in small samples. Monte Carlo results presented by Neykov et al. (2007) demonstrate that their robust discordancy method outperforms the measure recommended

by Hosking and Wallis (1997). Therefore, the robust discordancy measure will be implemented herein using the R package "rrcov" of Todorov and Filzmoser (2009).

2.2.1.2 Delineation of Homogeneous Regions

Multivariate statistical techniques such as cluster analysis, principal component analysis, canonical correlation analysis, and linear discriminant analysis are commonly employed to delineate homogeneous regions, i.e. group sites with similar extreme hydrologic response, and subsequently classify ungauged sites (see for example, Zrinji and Burn 1994; Burn 1997; Burn et al. 1997; Chiang et al. 2002a, 2002b; Rao and Srinivas 2006; Srinivas et al. 2008). Application of these methods requires the selection of appropriate similarity measures to characterize the extreme hydrologic response, or flood regime, at individual sites.

Possible similarity measures include at-site flood statistics, such as the magnitude of the T-year flood, the timing and/or duration of flood events, the conventional coefficient of variation (CV), and L-moment ratios (e.g., L-CV or L-Skewness). However, use of flood statistics to delineate regions may compromise the validity of the hydrological homogeneity test which is based on the same data (Burn 1990, 1997). As a result, the delineated regions may appear homogeneous, but would be inefficient for developing regional quantile estimators. In addition, unless the delineated regions are sufficiently contiguous, the successful classification of an ungauged basin within a delineated region cannot be guaranteed.

It is generally assumed that the flood regime, or hydrologic response, is similar at sites with comparable basin characteristics (e.g. physiographic characteristics and

meteorological inputs). (See for example, GREHYS 1996[a]; Hosking and Wallis 1997; Chiang et al. 2002a, 2002b; Isik and Singh 2008; Rao and Srinivas 2006.) Therefore, measurable characteristics such as the drainage area, land use and land cover metrics, geology, basin elevation and mean annual precipitation can be used in place of similarity measures based on gauged flood data. This is advantageous, as an ungauged site can then be classified within a delineated region, and regional quantile estimators can be developed for that site (Burn 1997; Burn et al. 1997).

Still, a major complication remains in that the regions formed are dependent on the basin characteristics employed as indicators of hydrologic similarity, as well as the statistical technique(s) used to delineate the regions (GREHYS 1996b; Castellarin et al. 2001). This concern will be addressed in Chapter 3; the following sections describe the multivariate statistical techniques employed therein.

Cluster Analysis

Cluster analysis (CA) groups sites on the basis of a statistical distance measure reflecting the similarity (or dissimilarity) among a set of attributes (similarity measures) selected to represent each gauging station. Several clustering techniques are available in the statistical literature, including hierarchical approaches such as single linkage, complete linkage, average linkage and Ward's method, as well as non-hierarchical approaches such as the k-means method (Johnson and Wichern 2007, p. 671). These methods have been widely used in the delineation of hydrologically homogeneous regions (see for example, Burn 1988, 1989, 1997, 2000; Bhaskar and O'Connor 1989; Baeriswyl and Rebetez 1997; Hosking and Wallis 1997; Chiang et al. 2002a; Castellarin

et al. 2001; Dinpashoh et al. 2004; Rao and Srinivas 2006). Ward's hierarchical method is most commonly used in the hydrologic literature as it tends to delineate regions roughly equivalent in size, and is thus considered more appropriate in the context of regionalizing flood data (Hosking and Wallis 1997, pp. 58-59). For these reasons, Ward's method will be employed herein.

Ward's method is an agglomerative hierarchical algorithm which initially begins with each site serving as its own cluster (or region). The algorithm successively merges clusters using an analysis of variance approach in which the similarity among members (or sites) in a region is measured in terms of the Error Sum of Squares (ESS). For region *k* containing N sites, wherein the flood regime is represented by p attributes (X_1, X_2, …, X_p), the ESS is given by

$$ESS_k = \sum_{j=1}^{N} \left(\boldsymbol{x}_j - \overline{\boldsymbol{x}}\right)' \left(\boldsymbol{x}_j - \overline{\boldsymbol{x}}\right) \tag{28}$$

where $\boldsymbol{x}_j = (x_1,\ x_2, \ldots, x_p)$ is a vector of the attributes at site j, and $\overline{\boldsymbol{x}} = \left(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_p\right)$ is a vector of the means of the attributes within the region. At each step, ESS_k is computed for the hypothetical merger of any two clusters, and the actual mergers chosen to occur are those which minimize the increase in the total ESS across all regions. A dendrogram is commonly used to illustrate the mergers made at successive levels, where the vertical axis represents the value of the ESS.

A pseudo F-test can be used to determine the number of regions (K) that should be delineated (Fovell and Fovell 1993). The test seeks to minimize the differences within each region (ESS), while maximizing the differences between the regions, termed the

Between Sum of Squares (BSS). In particular, the optimal number of regions to delineate (i.e., value of K) is that which maximizes the following function:

$$F = \frac{\mathrm{BSS}\left(\mathrm{N_T} - \mathrm{K}\right)}{\mathrm{ESS}\left(\mathrm{K} - 1\right)} \tag{29}$$

where N_T is the total number of sites and K is the current number of regions delineated. The Sum of Squares terms are computed using the following equations:

$$ESS = \sum_{k=1}^{K} ESS_k \tag{30}$$

$$BSS = \sum_{k=1}^{K} (\overline{\boldsymbol{x}}_k - \overline{\boldsymbol{x}})'(\overline{\boldsymbol{x}}_k - \overline{\boldsymbol{x}}) \tag{31}$$

where $\overline{\boldsymbol{x}}_k$ is a vector of the means of the p attributes within region *k*, and $\overline{\boldsymbol{x}}$ is a vector of the means of the p attributes computed across all of the basins regardless of the region delineation.

Cluster analysis using Ward's method is employed in Chapters 3 and 4 herein using a variety of basin characteristics as similarity measures, both as a lone procedure and in conjunction with output from principal component analyses and canonical correlation analyses. The latter procedures are discussed in the following sections.

Principal Component Analysis

Principal Component Analysis (PCA) is a multivariate statistical method that expresses the dataset in terms of uncorrelated linear functions of the original variables called principal components (PCs). The components are chosen so that they successively contain the maximum variability of the original dataset. When the first few PCs explain the majority of the variability, PCA identifies a fewer number of variables which can be

employed in subsequent analyses with little loss of information (Jolliffe 2002). Applications of PCA are found in a variety of fields including genetics (Yeung and Ruzzo 2001), remote sensing (Fung and Ledrew 1987), social sciences (Schyns 1998) and meteorology (Mallants and Feyen, 1990). PCA has also been employed to delineate regions for regional flood frequency analyses (e.g., Lins 1985, 1997; Detenbeck et al. 2005; Kahya et al. 2008).

The PCA methodology can be described as follows. Given a vector of p variables $\mathbf{X} = \left[X_1, X_2, ..., X_p \right]^T$ with a (p x p) variance-covariance matrix $\mathbf{\Sigma}_X$, the eigenvalues λ_i and eigenvectors $\mathbf{e}_i$ (for i = 1,…, p) associated with $\mathbf{\Sigma}_X$ can be estimated by solving the following equations:

$$\det\left(\Sigma_X - \lambda \mathbf{I}_p\right) = 0 \tag{32}$$

$$\left[\mathbf{\Sigma}_X\right]\mathbf{e}_i = \lambda_i \mathbf{e}_i \tag{33}$$

where $\mathbf{I}_p$ is a (p x p) identity matrix. The resulting eigenvalues are ordered so that $\lambda_1 > \lambda_2 > \ldots. > \lambda_p$. The eigenvectors specify the orthogonal directions of most variance, and their associated eigenvalues indicate the magnitude of the variance. Once the eigenvalues and eigenvectors of the variance-covariance matrix are obtained, the principal components (Y_i for i = 1,…,p) are defined as

$$\mathbf{Y} = \left[\mathbf{E}^{T}\right]\mathbf{X} \tag{34}$$

where $\mathbf{E} = \left[\{\mathbf{e}_1\},\{\mathbf{e}_2\},...,\{\mathbf{e}_p\}\right]$. If the correlation matrix $\mathbf{R}_X$ is used in place of the variance-covariance matrix $\mathbf{\Sigma}_X$, then the eigenvectors are of unit length and they convey the contribution of each original attribute (X_j) to a specific component (Y_i). Thus, the elements of the eigenvectors (e_{ij}) are called component loadings because a larger e_{ij} indicates greater importance of the corresponding variable X_j in component i. In addition, Y_i is the principal component score with variance equal to the corresponding eigenvalue λ_i, and all principal components are uncorrelated with one another.

$$\mathrm{Var}(\mathbf{Y}_i) = \mathrm{Var}\left(\mathbf{e}_i^{T}\mathbf{X}\right) = \mathbf{e}_i^{T}\left[\Sigma_X\right]\mathbf{e}_i = \lambda_i \tag{35}$$

$$\mathrm{Cov}(\mathbf{Y}_i,\mathbf{Y}_j) = \mathbf{e}_i^{T}\left[\mathbf{\Sigma}_X\right]\mathbf{e}_j = 0 \qquad \text{for } i \neq j \tag{36}$$

When dimension reduction is desirable, a number of components will be excluded as they do not contain a considerable amount of information (i.e., Y_{p-1}, Y_p). To decide on the number of components to exclude, one can use the scree plot of the cumulative variance in search of a plateau or, as a rule of thumb, retain the components that have a variance (or eigenvalue, λ_i) greater than 1 when the correlation matrix is used (Johnson and Wichern 2007, p. 451). In the research presented herein, the first few principal components (with $\lambda_i > 1$) should contain the basin characteristics that explain most of the variability among the sites (and corresponding watersheds) in the study area, and thus would be useful in delineating homogeneous regions. The first few PCs will also define new variables which will subsequently be used in cluster analyses.

Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) provides an estimate of the existing association between two sets of variables (Johnson and Wichern 2007, p. 539). Linear combinations of the variables in one set are paired with linear combinations of the variables in the second set to form canonical variables. The first canonical variable is the pair of linear combinations with unit variance and the maximum correlation. Subsequent canonical variables with unit variance are formed such that the correlation is again maximized while ensuring all canonical variables are uncorrelated. The correlation for a given pair of canonical variables is called the canonical correlation. Only a few studies have employed CCA to delineate hydrologically homogeneous regions for the regionalization of flood data (Ribeiro-Correa et al. 1995; Ouarda et al. 2001; Hache et al. 2002).

The CCA methodology can be described as follows. Consider two sets of variables represented by the random normalized vectors $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_q)^T$ with means μ_X and μ_Y and correlation matrices $\mathbf{R}_{XX}$ and $\mathbf{R}_{YY}$, respectively. In addition, the correlation between **X** and **Y** is represented by the (p x q) matrix $\mathbf{R}_{XY}$ where $p \leq q$. In the context of delineating homogeneous regions, **X** consists of hydrological variables, and **Y** consists of the measurable watershed characteristics. We define linear combinations of the variables in **X** and **Y** such as $U = \mathbf{a}^T\mathbf{X}$ and $V = \mathbf{b}^T\mathbf{Y}$. U and V are canonical variates if the vectors **a** and **b** successively maximize the correlation between the pairs (U_i, V_i) for $i = 1, \ldots, p$ and the correlation of (U_i, V_i) to any

other pair (U_j, V_j) for $i \neq j$ is zero. The canonical vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ which define the i^{th} pair of canonical variables are computed as follows:

$$\mathbf{a}_i = \mathbf{R}_{XX}^{-1/2}\mathbf{e}_i \tag{37}$$

$$\mathbf{b}_i = \mathbf{R}_{YY}^{-1/2}\mathbf{f}_i \tag{38}$$

where $\mathbf{e}_i$ represents the eigenvectors associated with $\mathbf{R}_{XX}^{-1/2}\mathbf{R}_{XY}\mathbf{R}_{YY}^{-1}\mathbf{R}_{YX}\mathbf{R}_{XX}^{-1/2}$, and $\mathbf{f}_i \propto \mathbf{R}_{YY}^{-1/2}\mathbf{R}_{YX}\mathbf{R}_{XX}^{-1/2}\mathbf{e}_i$. In order for $\mathbf{f}_i$ to be uniquely specified, the following constraint of unit variance must be satisfied: $\text{var}(\mathbf{b}_i^T\mathbf{R}_{YY}\mathbf{b}_i) = 1$. The individual coefficients within $\mathbf{a}_i$ and $\mathbf{b}_i$ represent the relative importance of specific variables within **X** and **Y** in the linear combination.

The canonical correlation between the two variables U and V is calculated using:

$$\text{Corr}(U, V) = \frac{\mathbf{a}^T\mathbf{R}_{XY}\mathbf{b}^T}{\sqrt{\mathbf{a}^T\mathbf{R}_{XX}\mathbf{a}}\sqrt{\mathbf{b}^T\mathbf{R}_{YY}\mathbf{b}}} \tag{39}$$

In the research presented herein, the first few canonical variables with the highest correlation should indicate which basin characteristics would serve as useful indicators of hydrological homogeneity for use in region delineation.

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a multivariate technique used to identify the variables that are best able to explain differences between a set of predefined regions. LDA yields linear functions of the original variables, termed discriminant functions. The first few discriminant functions best discriminate between the regions, and therefore, can

be used for classification of a future dataset. In the context of regionalization of hydrologic data, these discriminant functions would be useful for allocating an ungauged site to a delineated region. However, few studies in the hydrologic literature have employed LDA (Chiang et al. 2002a, 2002b; Detenbeck 2005; Cianfrani et al. 2006; Snelder et al. 2009), and none have used it in the context of flood frequency analysis.

The LDA methodology can be described as follows. Consider K groups (or regions) composed of watersheds characterized by p variables. The (p x 1) vectors $\overline{\boldsymbol{x}}_1$, $\overline{\boldsymbol{x}}_2$, ... $\overline{\boldsymbol{x}}_K$ contain the means of each variable within region k (k = 1, 2, … K), and the (p x 1) vector $\overline{\boldsymbol{x}}$ contains the overall means for each of the p variables computed across all K regions. It is assumed that the K regions are multivariate normal and that their (p x p) variance-covariance matrices are all equal to **S**. If not, **S** should be replaced by the pooled sample variance-covariance matrix $\mathbf{S}_{\text{pooled}}$ computed as:

$$\boldsymbol{S}_{\text{pooled}} = \frac{1}{N_1+N_2+\cdots+N_K}\left((N_1-1)\boldsymbol{S}_1 + (N_2-1)\boldsymbol{S}_2 + \cdots + (N_K-1)\boldsymbol{S}_K\right) \qquad (40)$$

where $\mathbf{S}_k$ is the variance-covariance matrix of region k (k = 1, 2, … K) containing N_k sites.

The discriminant functions are chosen to maximize the differences between regions. This requires defining the sample between groups sum of cross products matrix (**B**) to measure the variation of each of the regional means ($\overline{\boldsymbol{x}}_k$) around the overall mean ($\overline{\boldsymbol{x}}$) for each variable, and the sample within groups matrix (**W**) to measure the variation of variables within a region about the regional mean ($\overline{\boldsymbol{x}}_k$). These matrices are computed as follows:

$$\mathbf{B} = \sum_{k=1}^{K} (\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}})^T \tag{41}$$

$$\mathbf{W} = \sum_{k=1}^{K} \sum_{j=1}^{Nk} \left(x_{jk} - \bar{\boldsymbol{x}}_k\right)\left(x_{jk} - \bar{\boldsymbol{x}}_k\right)^T \tag{42}$$

wherein N_k is the number of sites (or watersheds) in region k, and x_{jk} is the j^{th} site in region k. The eigenvalues λ_i and eigenvectors $\mathbf{e}_i$ [for i = 1,…,s where s = min(K-1, p)] associated with the matrix $\mathbf{W}^{-1}\mathbf{B}$ are then estimated by solving the following equations:

$$|\mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I}| = 0$$

$$(\mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I})\boldsymbol{e}_i = 0 \tag{43}$$

The resulting eigenvalues are ordered so that $\lambda_1 > \lambda_2 > \ldots. > \lambda_s$ and the eigenvectors are normalized so that $\boldsymbol{e}_i^T \boldsymbol{S}_{pooled} \boldsymbol{e}_i = 1$. The discriminant functions (Y_i for i = 1, …, s) are then defined as:

$$\mathbf{Y} = \left[\mathbf{E}^T\right]\mathbf{X} \tag{44}$$

where $\mathbf{E} = \left[\{\mathbf{e}_1\},\{\mathbf{e}_2\},...,\{\mathbf{e}_s\}\right]$ and $\mathbf{X} = (X_1, X_2, \ldots., X_p)^T$ contains the p watershed variables. The magnitudes of the eigenvalues convey the power of the discriminant function to differentiate between regions. In the research presented in Chapter 3, the first few discriminant functions will indicate which basin characteristics should be used for delineation of hydrologically homogeneous regions. In order to develop quantile estimators at ungauged locations, discriminant functions will also be used in Chapter 5 to classify ungauged sites within delineated regions.

2.2.2 Generalized Least Squares Regression

Generalized least squares (GLS) regression has been used extensively throughout the U.S. and the world to estimate flood statistics at ungauged sites as a function of physiographic characteristics of a watershed such as drainage area, main-channel slope, average annual precipitation, and land use and land cover indices (e.g., Tasker and Stedinger 1989; Madsen and Rosberg 1997; Griffis and Stedinger 2007a). Stedinger and Tasker (1985, 1986a, 1986b) and Kroll and Stedinger (1998) show that GLS estimators are much more appropriate and efficient for use with hydrologic data than ordinary least squares (OLS) estimators. Unlike OLS estimators, the GLS estimators account for differences in the variance of streamflows from site-to-site due to different record lengths, and cross-correlation of the estimators due to correlation among concurrent streamflows at nearby sites (Tasker 1980; Kuczera 1983).

Applications of GLS regression with the GEV distribution often involve estimating the at-site L-moments (λ_1, τ_2, and τ_3) as a function of watershed or climate characteristics using log-log or log-linear relationships (e.g., Reis 2005; Jeong et al. 2007). For example, the first L-moment (λ_1) can be related to a set of k basin characteristics using the following expression:

$$\log_{10}\left(\lambda_1^{(i)}\right) = \beta_o + \sum_{j=1}^{k} \beta_j X_{ij} + \delta_i \tag{45}$$

where X_{ij} is the base-10 logarithm of the j^{th} basin characteristic corresponding to site i, β_j is the corresponding regression coefficient, and δ_i is the model error. When equation (45) is applied to N sites in a region (i = 1, …, N), the resulting model errors are assumed

to be independent and normally distributed with mean zero and variance σ_δ^2. Similar expressions could be used to represent higher order L-moments or L-moment ratios (e.g., L-CV or L-Skewness).

Typically, regression coefficients are obtained assuming the values of the predictor variables, and the corresponding response variables, are measured with negligible error. In hydrologic applications, however, the true value of the response variable (e.g., λ_1) is unknown and must be estimated using an at-site analysis for sites with gauged data. A time-sampling error, η_i, results due to limited record lengths. In general, this error can be expressed as

$$\hat{y}_i = y_i + \eta_i \tag{46}$$

where 10^{y_i} is the true value of the L-moment in question (e.g. λ_1, τ_2, or τ_3) at site i, and $10^{\hat{y}_i}$ is an estimate of the L-moment using the available at-site data. Given that L-moment estimators are unbiased estimators of the population L-moments, η_i has a mean of zero; the variance of η_i is a function of the estimated sample moments. For example, the sampling error variance associated with the logarithm of the first L-moment is given by (Madsen and Rosbjerg 1997):

$$Var\left[\log_{10}\left(\hat{\lambda}_1\right)\right] = \frac{CV_i^2}{n_i} \tag{47}$$

where n_i is the record length available at site i, and CV_i is the population coefficient of variation at site i.

For a regional analysis across N sites, equations (45) and (46) are combined resulting in the following relationship in matrix notation

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{48}$$

where **Y** is a (N x 1) vector of the base-10 logarithms of the sample L-moments, **X** is a [N x (k+1)] matrix of the base-10 logarithms of the basin characteristics with an added column of ones corresponding to the intercept, $\boldsymbol{\beta}$ is a [(k+1) x 1] vector containing the regression coefficients, and $\boldsymbol{\epsilon} = \boldsymbol{\eta} + \boldsymbol{\delta}$ is a vector of random errors with $E[\boldsymbol{\epsilon}] = 0$ and $E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \boldsymbol{\Lambda}$. In this way, the variance of residuals has two components: (i) the model error variance σ_δ^2, which is a measure of the precision with which the true model can predict L-moments, and (ii) the sampling-covariance matrix **Σ** which accounts for the time-sampling variance in the logarithms of the at-site L-moment estimators, as well as cross-correlation with estimators at nearby sites due to correlation among concurrent streamflows (e.g., Tasker and Stedinger 1989; Madsen et al. 2002; Martins and Stedinger 2002).

The coefficients of the regression model and the model error variance are obtained using an iterative procedure to solve the following equations (Stedinger and Tasker 1985):

$$\widehat{\boldsymbol{\beta}}_{GLS} = \left(\boldsymbol{X}^T\widehat{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{\Lambda}}^{-1}\widehat{\boldsymbol{Y}} \tag{49}$$

and

$$\widehat{\boldsymbol{\Lambda}}\left(\sigma_\delta^2\right) = \sigma_\delta^2\boldsymbol{I}_{\boldsymbol{N}} + \widehat{\boldsymbol{\Sigma}} \tag{50}$$

where $\boldsymbol{I}_{\boldsymbol{N}}$ is an (N x N) identity matrix and elements of the sampling covariance matrix are estimated using available streamflow data. For example, the estimator of **Σ** for the logarithm of the first L-moment is computed as (Madsen and Rosbjerg 1997):

$$\hat{\Sigma}_{ii} = \frac{\widehat{CV}_i^2}{n_i} \qquad \text{for i = j}$$

$$\hat{\Sigma}_{ij} = \frac{\rho_{ij} m_{ij} \widehat{CV}_i \widehat{CV}_j}{n_i n_j} \qquad \text{for i} \neq \text{j} \tag{51}$$

where $\widehat{CV}_i$ and $\widehat{CV}_j$ are the sample coefficients of variation at sites i and j, respectively, n_i is the record length at site i, n_j the record length at site j, m_{ij} is the number of concurrent years of record at sites i and j, and ρ_{ij} is the lag zero cross-correlation of flows between sites i and j. To avoid correlation among the residuals, the sample coefficients of variation are estimated as a function of drainage area using a separate regional regression, and ρ_{ij} is computed as a function of the distance between sites i and j as recommended by Tasker and Stedinger (1989). GLS regression models for the L-CV and L-Skewness can be obtained in a similar fashion; however, mathematical expressions for the necessary sampling covariance matrices are unavailable. Martins and Stedinger (2002) describe the use of Monte Carlo simulations to infer the values of those expressions.

As multiple models can be created using various combinations of basin characteristics, statistical measures are used to avoid over parameterization of the model and to ensure selection of the most relevant explanatory variables. Standard metrics for the assessment of GLS regression model precision in hydrologic applications include the model error variance, the average variance of prediction (AVP), and the pseudo-R^2 (e.g., Griffis and Stedinger, 2007a). The latter is computed as

$$R_{GLS}^2 = 1 - \frac{\hat{\sigma}_\delta^2(k)}{\hat{\sigma}_\delta^2(0)} \tag{52}$$

where $\hat{\sigma}_\delta^2(k)$ is the model error variance for the regression model with k explanatory variables, and $\hat{\sigma}_\delta^2(0)$ is the estimated model error variance when no explanatory variables are employed, i.e. only the constant β_0 is retained. This pseudo-R^2 is preferred over the traditional R^2 and adjusted-R^2 which misrepresent the true power of the GLS model as they measure the proportion of variance in the observed values of the response variable (e.g. L-moment estimates) explained by the fitted model. Unfortunately, that proportion considers the total error, which includes the sampling error.

Under the assumption that the regional model was developed using data collected at sites which are representative of those at which predictions will be made, the average variance of prediction (AVP) over the available dataset is a measure of how well the GLS regression model predicts the true L-moment on average (Tasker and Stedinger 1986). The AVP is computed as

$$AVP_{GLS} = \sigma_\delta^2 + \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \left(\boldsymbol{X}^T \hat{\boldsymbol{\Lambda}}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{x}_i^T \tag{53}$$

where $\boldsymbol{x}_i$ is a row vector containing the logarithms of the basin characteristics for site i. When comparing hydrologic regression models a smaller AVP is preferred. (See for example, Reis et al. 2005; Griffis and Stedinger 2007a, and citations therein.)

In the research presented in Chapter 5 herein, relationships between the index-flood parameters and basin characteristics are derived using GLS regression analyses. These relationships will subsequently be used to predict flow quantiles at ungauged sites. Preference will be given to models with the smallest σ_δ^2 and AVP, and the largest R_{GLS}^2.

Chapter 3 : A Standardized Procedure for Delineation of Hydrologically Homogeneous Regions

Application of the Index Flood procedure at ungauged sites requires that hydrologically homogeneous regions be delineated using physical similarity measures in lieu of flood statistics. Physical characteristics can then be used to classify an ungauged site within a delineated region, and subsequently estimate needed flood quantiles. Unfortunately, the delineated regions are highly dependent on the choice of similarity measures used to characterize the extreme hydrologic response at individual sites, as well as the statistical techniques employed to group similar sites (see for example, GREHYS 1996b; Burn 1990, 1997; Burn et al. 1997; Hosking and Wallis 1997; Castellarin et al. 2001). In addition, once regions are obtained, the identification of and decision to remove discordant sites is highly subjective. Typically, efforts to improve regional homogeneity as measured by the H-statistic involve moving hydrologically discordant sites from one region to another without any clearly defined physical basis (Rao and Srinivas 2006; Srinivas et al. 2008). These are major limitations of the Index Flood method, and significantly impair its ability to provide accurate quantile estimates at ungauged sites.

The goal of this chapter is to decrease the subjectivity associated with the delineation of hydrologically homogeneous regions for regional flood frequency analyses. Herein, a standardized procedure for region delineation is proposed and evaluated using data for sites across the Southeastern United States. Key components of this procedure are a new statistical metric to identify physically discordant sites and a new methodology to identify the physical variables that are the most indicative of

extreme hydrologic response. In addition, results presented herein are used to validate the hypothesis that the identified physical attributes can be used to infer the flood regime and successfully estimate quantiles in areas outside of the extent of the study area used for model development.

3.1 Standardized Procedure for Region Delineation

A new procedure for attribute selection and region delineation is outlined below. The procedure employs techniques commonly employed in practice, but offers simple rules and guidelines to remove subjectivity from the delineation process while ensuring the best possible classification of ungauged sites in homogeneous regions. An application of this process is presented in subsequent sections to provide a more detailed explanation of key steps.

As is typical in practice, the process begins with compiling a set of non-redundant physical variables relevant to the flood regime. Usually, all available variables are then used as similarity measures in multivariate statistical techniques to delineate regions, without prior consideration of their impact on the flood regime. We recommend an intermediate step wherein the physical variables most relevant to the flood regime are identified using a combination of cluster analyses, principal component analyses, canonical correlation analyses and linear discriminant analyses applied to sets of flood statistics and physical variables. Although many studies delineate regions based on physical characteristics, efforts to select an appropriate set of characteristics for region delineation are rarely made in practice (Bates 1997). However, as demonstrated later in

this chapter, doing so results in improving the hydrologic homogeneity of delineated regions, and thus, more efficient quantile estimators for use at ungauged sites are obtained.

The physical variables identified as being most indicative of the flood regime are used as attributes in a cluster analysis using Ward's method (see Section 2.2.1.2) to delineate regions suitable for regional flood frequency analysis. Regions are formed such that the Error Sum of Squares across all regions is minimized, while ensuring that each region contains at least 10 sites. Given the standard rule of thumb that 5T years of record are needed within a region to accurately estimate the T-year flood, and assuming that the average record length at an individual site is 50 years, a minimum of 10 sites within each region ensures that enough data is available to estimate the 100-year event at ungauged sites.

The hydrological homogeneity of each delineated region is then assessed using the H-statistic of Hosking and Wallis (1997) calculated using equation (23). If $H < 4$, then the region is sufficiently homogeneous and can be used to generate flood quantile estimates at ungauged sites; if $H \geq 4$, then discordant sites should be identified and removed from the region in order to improve the homogeneity (Hosking and Wallis 1993; Robson and Reed 1999; Guse et al. 2009). Sites which are hydrologically discordant are identified using the robust discordancy test calculated using equation (26) as proposed by Neykov et al. (2007). Typically, any sites which are identified as hydrologically discordant are removed without consideration of their physical characteristics. This is contrary to the recommendations of Hosking and Wallis (1997) who advocate that only

those sites which are both hydrologically and physically discordant be removed from a delineated region. However, no metrics to assess the physical discordancy of sites have previously been provided. Herein, we propose a new metric to evaluate the physical discordancy of sites and their corresponding watersheds (see below). Only those sites that are both hydrologically and physically discordant should be removed from regions with $H \geq 4$. If discordant sites are removed and the H value remains greater than 4, then quantile estimates derived using the Index Flood method should be used with caution.

3.1.1 Physical Discordancy Test

Physically discordant watersheds are markedly different from others within a given region with regards to their basin characteristics. Using data for the N sites within the region in question, physically discordant watersheds (and corresponding sites) can be identified using output from a principal component analysis performed on all available physical variables which have been appropriately normalized and centered by subtracting their mean values. All PCs with standard deviation (or eigenvalues, λ_i) greater than one should be retained as these are the components which explain most of the variability within the region (see Johnson and Wichern 2007, p. 451, and Section 2.2.1.2 herein), and therefore, will be most useful in identifying discordant sites. Assuming that the PC scores of the retained components follow a normal distribution, a critical T-value can be computed as follows:

$$T_{critical} = \pm t_{0.975,N-1} S_i \tag{54}$$

where $t_{0.975,N-1}$ is the student t value with N-1 degrees of freedom corresponding to 0.975 cumulative probability, and $S_i = \sqrt{\lambda_i}$ is the standard deviation of the principal component under examination. This is equivalent to constructing a 95% confidence interval for the mean when the mean of the principal component is zero, as is the case in PCA calculations when the observations are centered by subtracting their mean. Watersheds with PC scores above or below the critical value are deemed physically discordant. Appendix B provides an example which illustrates the application of this test for physical discordancy.

3.2 Data and Study Location

The study area spans the states of Georgia, South Carolina, and North Carolina, as well as neighboring portions of Alabama, Florida, Tennessee, and Virginia. A total of 249 unimpaired and unnested watersheds and their corresponding U.S. Geological Survey (USGS) gauging stations are considered. The locations of these gauging stations are illustrated in Figure 3.1. These sites were identified using the Hydro-Climatic Data Network (Slack et al. 1993). Annual maximum instantaneous peak flow data were collected for each site from the USGS website (http://nwis.waterdata.usgs.gov/usa/nwis/peak). Record length statistics for the flood series compiled by state are presented in Table 3.1. Overall, the gauging stations have relatively long records with a maximum of 115 years, an average on the order of 50 years, and a minimum of 16 years.

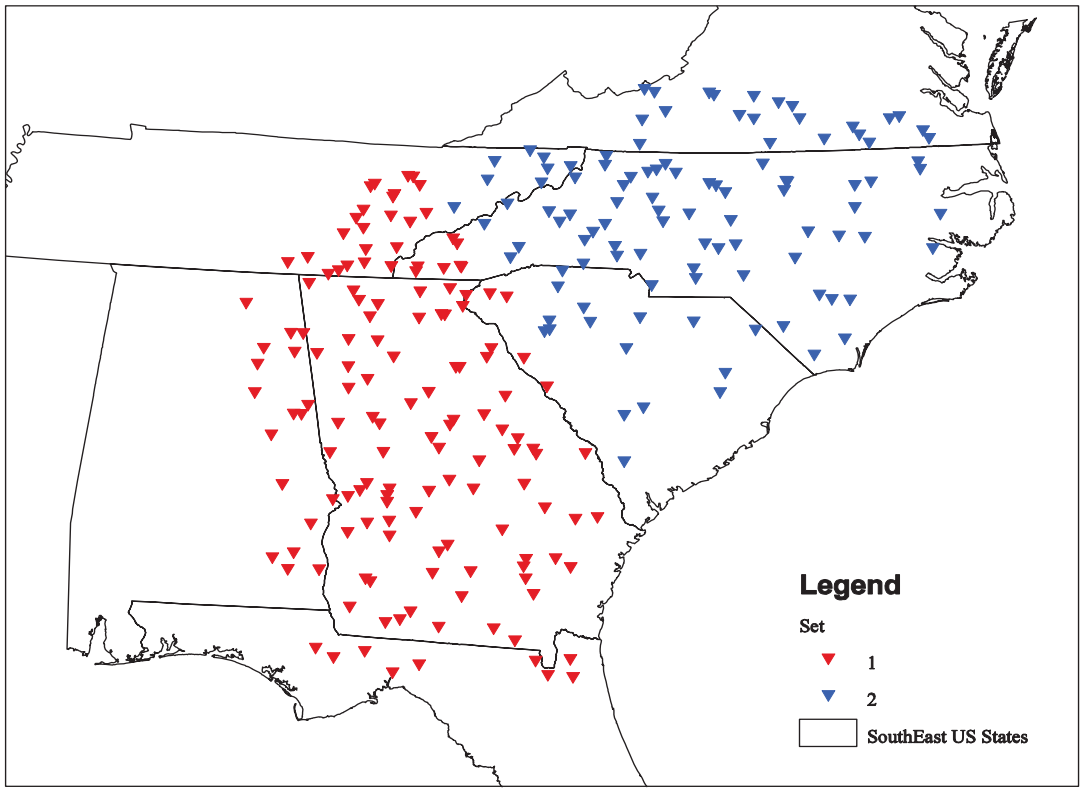


Figure 3.1: Location of unimpaired and unnested gauging stations within the Southeastern U.S. Sites in Set 1 comprise the training set; sites in Set 2 comprise the validation set. (State boundary files downloaded from Census Website: http://www.census.gov/geo/www/cob/st2000.html)

Table 3.1
Number of sites per state and record length statistics.

State	Number of sites	Minimum record length	Maximum record length	Average record length
Alabama	18	25	77	49
Florida	9	27	72	44
Georgia	82	25	115	41
N. Carolina	67	27	107	56
S. Carolina	20	27	79	53
Tennessee	31	16	78	44
Virginia	22	25	108	56

In order to demonstrate the efficiency of the proposed standardized method in alternate study areas, the 249 sites are divided into two sets: Set 1 composed of 143 sites constitutes the training set located in the western portion of the Southeastern U.S. (represented by red triangles in Figure 3.1); Set 2 composed of 106 sites constitutes the validation set located in the eastern portion of the Southeastern U.S. (represented by blue triangles in Figure 3.1). Although a similar flood regime is not guaranteed by geographic proximity, a geographic division rather than a random division of the dataset is necessary for the applications herein. Data for sites contained in Set 1 will be used in subsequent sections of this chapter to illustrate the application of the standardized method for region delineation, and in jackknife resampling procedures used to evaluate the efficiency of the delineated regions for quantile estimation at ungauged sites. Data for sites contained in Set 2 will be used to test the hypothesis that the attributes selected for Set 1 will be appropriate to characterize the flood regime in Set 2. Verification of this hypothesis is necessary for the applications presented in Chapter 5 for ungauged basins beyond the extent of the study area.

A variety of hydrologic and physical parameters were obtained for each of the 249 sites. Using available at-site flow data, the first three L-moment ratios (L-CV, L-Skewness, and L-Kurtosis) were calculated for each site, and the 10-year, 25-year, 50-year, and 100-year flood quantiles were estimated using the GEV distribution with parameters estimated by the L-moments method (see Chapter 2.1.2). These L-moment ratios and flood quantiles are used to represent the flood regime in subsequent analyses. In addition, the following nine basin characteristics were collected for each site: drainage

area (A) in square miles, main channel slope (S_{Ch}) in feet/mile, mean basin slope (S_B) in percent, basin shape (S_h) computed as the ratio of the main channel length squared to the drainage area (dimensionless), mean basin elevation (E) in feet, percent impervious surface (Imp), percent forest cover (F), a soil drainage index (SI) used to approximate the drainage class of the watershed, and an infiltration index (Inf) used to approximate the infiltration rate in the basin. Details of the soil drainage and infiltration indices are provided in Appendix A. These nine physical characteristics represent variables typically considered in regional flood frequency analyses. Values of the physical characteristics for each basin under consideration herein were provided by Gruber and Stedinger (2008).

Summary statistics for each physical variable compiled by set are reported in Table 3.2. Except for main channel slope, the degree of variability in each physical parameter is similar in both datasets. However, two-sample hypothesis tests performed on the difference in the means of each physical variable in Set 1 versus Set 2 reveal that, with the exception of drainage area, the characteristics comprising the watersheds within the two datasets are significantly different at the 5% level. This suggests that the chosen datasets are sufficient to demonstrate application of the proposed standardized method for region delineation in alternate areas, and to assess the use of attributes selected for Set 1 to characterize the flood regime in Set 2.

Table 3.2
Summary statistics for the physical variables in Sets 1 and 2 of the study area.

Physical Characteristic	Minimum	Maximum	Average	Standard Deviation
		Set 1		
Drainage Area	9.60	2914	226.8	316.6
Main Channel Slope	0.60	737.5	41.1	105.2
Basin Slope	0	50.7	10.2	10.8
Basin Shape	2.40	28.3	6.90	4.30
Basin Elevation	36.3	4057	811.8	824.4
% Impervious Cover	0	40.5	4.30	6.00
% Forest Cover	0	51.8	99.2	24.4
Soil Drainage Index	2.30	6.20	3.40	0.80
Infiltration Index	1.90	2.40	3.80	0.40
		Set 2		
Drainage Area	2.72	2587	249	355.6
Main Channel Slope	0.71	189	17.6	26.7
Basin Slope	0.30	46.8	14.1	12.5
Basin Shape	1.60	30.7	8.50	4.90
Basin Elevation	26.6	3979	1076.3	1069.2
% Impervious Cover	0.02	34.3	2.14	4.30
% Forest Cover	8.32	99.0	56.6	20.0
Soil Drainage Index	2.30	3.35	5.90	0.70
Infiltration Index	1.90	3.90	2.40	0.40

Although the multivariate statistical procedures used herein do not explicitly assume normality, goodness-of-fit tests used to estimate their adequacy do. Therefore, the 249 data points corresponding to each physical variable were tested for normality using the probability plot correlation test (Stedinger et al. 1993), and efforts were made to normalize the data using Box-Cox transformations as necessary. For a given set of data denoted by x, the Box-Cox transformation searches for an exponent λ that creates a new

variable y which is approximately normally distributed (Box and Cox 1964). The new variable (y) represents a transformation of the original variable (x), and is obtained as follows:

$$y = \begin{cases} \frac{x^{\lambda}-1}{\lambda} \; if \;\; \lambda \neq 0 \\ \ln(x) \;\; if \;\; \lambda = 0 \end{cases} \tag{55}$$

Equation (55) was used to apply the Box-Cox transformation to each of the physical variables considered herein resulting in the values of lambda reported in Table 3.3. Summary statistics for the transformed variables compiled by Set are presented in Table 3.4. The transformed variables are subsequently standardized by scaling the values from zero to one to avoid giving more weight to any one variable. These normalized and standardized variables will be used in the applications of multivariate statistical techniques to follow; however, according to standard measures of normality, only the drainage area, channel slope and basin shape were successfully transformed.

Table 3.3
Exponent for Box-Cox transformations of physical variables.

Physical Characteristic	**λ**
Drainage Area	0
Main Channel Slope	-0.2
Basin Slope	0.3
Basin Shape	-0.3
Basin Elevation	0.1
% Impervious Cover	0.2
% Forest Cover	0.8
Soil Drainage Index	-3.0
Infiltration Index	-3.3

Table 3.4
Summary statistics of the transformed (normalized) physical variables in Sets 1 and 2.

Physical Characteristic	Minimum	Maximum	Average	Standard Deviation
		Set 1		
Drainage Area	0.98	3.46	2.12	0.47
Main Channel Slope	-0.54	3.67	1.81	0.77
Basin Slope	-3.23	7.49	2.55	2.21
Basin Shape	0.77	2.11	1.37	0.27
Basin Elevation	4.32	13.0	8.62	2.09
% Impervious Cover	-4.21	5.48	0.63	2.15
% Forest Cover	-1.25	48.2	27.4	12.1
Soil Drainage Index	0.305	0.331	0.323	0.004
Infiltration Index	0.266	0.299	0.284	0.007
		Set 2		
Drainage Area	0.43	3.41	2.07	0.57
Main Channel Slope	-0.35	3.25	1.74	0.66
Basin Slope	-1.01	7.23	3.41	1.99
Basin Shape	0.44	2.14	1.48	0.29
Basin Elevation	3.88	12.9	9.00	2.40
% Impervious Cover	-2.71	5.14	0.16	1.24
% Forest Cover	5.56	48.1	30.0	9.09
Soil Drainage Index	0.305	0.331	0.322	0.004
Infiltration Index	0.266	0.299	0.284	0.006

3.3 Attribute Selection

In this section, the physical variables that are most indicative of extreme hydrologic response in the western portion of the Southeastern U.S. (Set 1) will be identified by comparing and contrasting the characteristics of regions delineated based on flood statistics to those of regions based on physical variables. From here on out, these regions will be referred to as Statistics-Based Regions and Physically-Based Regions, respectively. Although flood statistics should not be used to delineate regions, the Statistics-Based Regions will represent the best case scenario for hydrologic similarity. These analyses will employ the normalized and standardized values of the nine physical variables for the sites in Set 1.

3.3.1 Statistics-Based Regions

Using the data in Set 1, Statistics-Based Regions are created by employing the L-CV as a similarity measure in the Ward clustering technique. The dendrogram in Figure 3.2 shows the distance or ESS between each cluster at each level of the algorithm. At the bottom of the figure, corresponding to a distance of 0, each vertical line represents an individual site. (Due to the number of sites are considered in the analysis, the actual station IDs corresponding to each line cannot be provided in the figure.) Aside from using the L-CV as a similarity measure, the Statistics-Based Regions are created following the guidelines outlined above in Section 3.1. Therefore, three regions are delineated to avoid the creation of clusters containing less than 10 sites. The spatial partition of the three Statistics-Based Regions is shown in Figure 3.3. It is clearly evident

that the regions are not contiguous, and thus it would be difficult to classify ungauged sites within one of the delineated regions.

For each Statistics-Based Region, Table 3.5 reports the number of sites therein (N), the heterogeneity measure (H), and the total record length available ($\sum_{all\ i} n_i$). These metrics are presented for the original regions resulting from the Ward clustering technique, as well as for regions modified by removing sites identified as hydrologically discordant (HD), and for regions modified by removing only the sites which are both hydrologically and physically discordant (HPD). The discordant sites are also noted in Figure 3.3, wherein circles denote HD sites and squares denote HPD sites; the color of the symbol corresponds to the original region within which the site in question was classified. The original regions are all sufficiently homogeneous with $H < 4$, and thus procedures to identify and remove discordant sites are unnecessary. However, as the results in Table 3.5 reveal, the homogeneity of the regions can be further improved if HD sites are removed. But, this is ill-advised in the interest of applications with ungauged sites; only sites which are HPD should be removed to ensure that the regional distribution would appropriately scale to any ungauged site allocated therein using physical characteristics. Both Regions 1 and 3 contain an individual HPD site, however, removal of the HPD site from Region 1 actually increases the statistical heterogeneity of the region. While Region 1 remains sufficiently homogeneous following removal of the HPD site, the increased heterogeneity of the region suggests that removal of discordant sites is in fact unnecessary when $H < 4$.

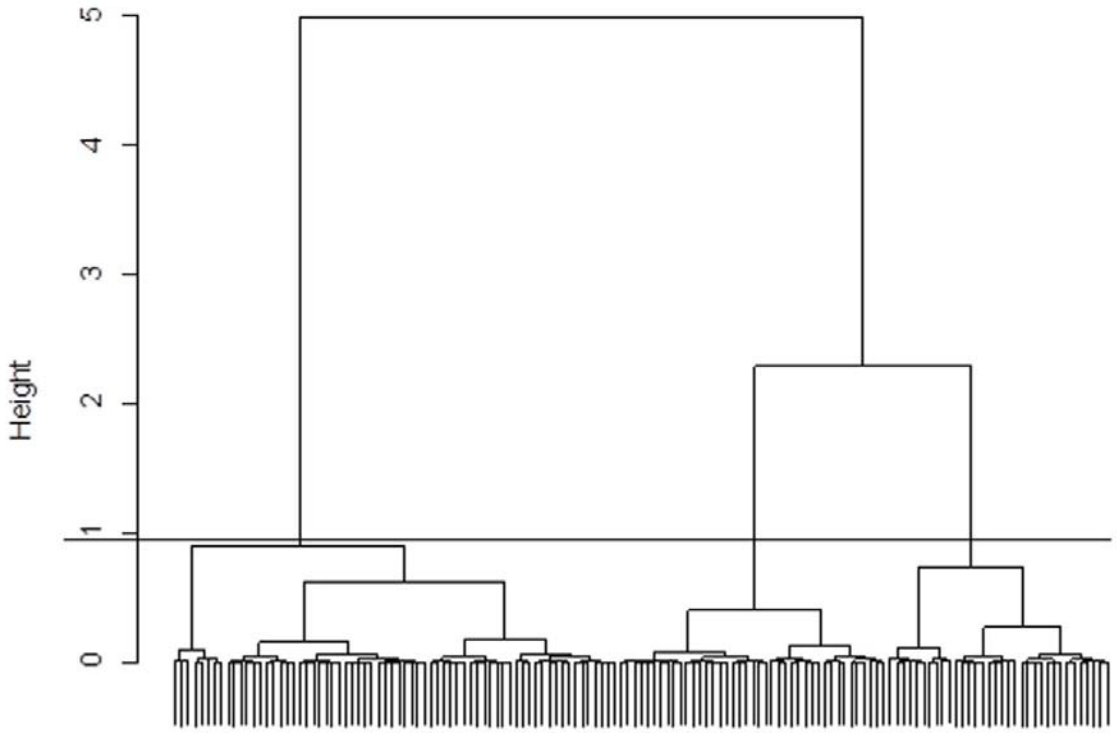


Figure 3.2: Dendrogram for Wards clustering applied to at-site L-CVs in Set 1.

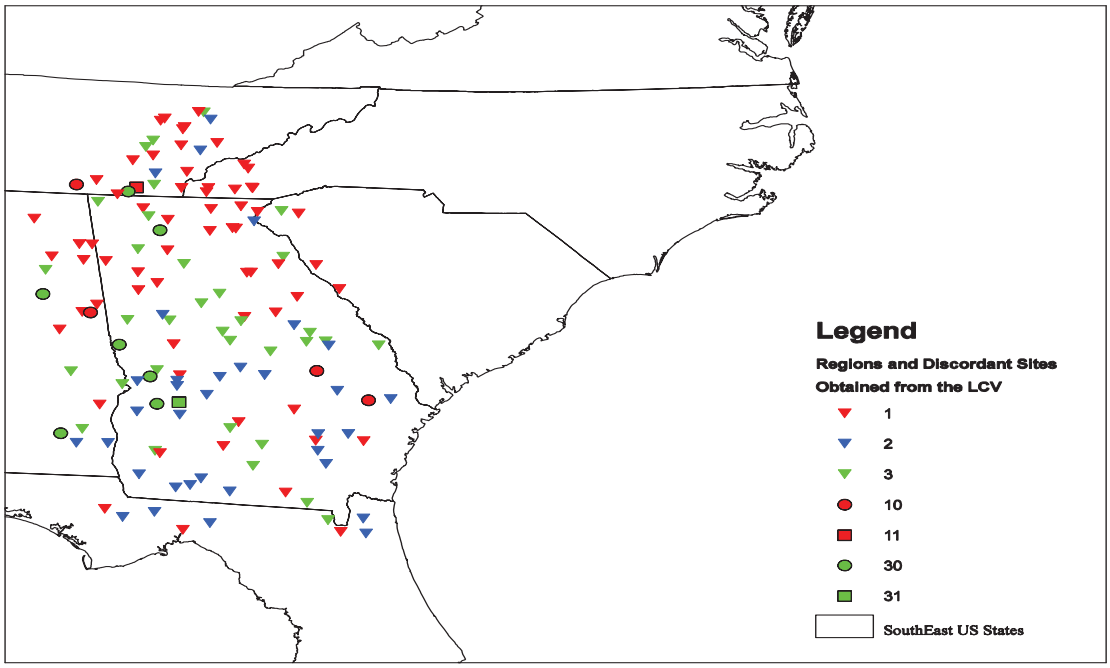


Figure 3.3: Three Statistics-Based Regions delineated using Wards clustering applied to at-site L-CVs in Set 1. Circles represent HD sites, squares represent HPD sites.

Table 3.5
Size and homogeneity of Statistics-Based Regions delineated using Wards clustering applied to at-site L-CVs in Set 1.

	Original Regions			**Regions without HD sites**			**Regions without HPD sites**		
	N	H	$\sum_{all\ i} n_i$	N	H	$\sum_{all\ i} n_i$	N	H	$\sum_{all\ i} n_i$
Region 1	68	1.28	3534	62	0.14	3322	67	1.51	3503
Region 2	34	-2.57	1285	33	-2.59	1252	34	-2.57	1285
Region 3	41	-5.11	1716	34	-5.34	1419	40	-5.28	1588

A linear discriminant analysis applied using the physical attributes of the Statistics-Based Regions will identify the physical variables that explain most of the differences between the homogeneous regions. The values of the nine physical variables (normalized and standardized) for the sites contained in each of the Statistics-Based Regions define three sets of variables. As there are only three sets, two linear discriminant functions are sufficient to differentiate between the regions. Table 3.6 reports the correlation of each physical variable to the linear discriminant functions. The first linear discriminant function (LD_1) explains 82% of the differences, and is mostly impacted by the basin slope, basin elevation, soil drainage and forest cover. It is therefore likely that those four physical characteristics would be useful in determining hydrological homogeneity with respect to flood flows in this area. The coefficients of LD_2 indicate that percent impervious cover is also influential, and may be a key indicator of the flashiness of the response from a given watershed. It is not considered in

subsequent steps to delineate regions, however, as LD_2 explains less than 20% of the differences between the regions.

Table 3.6
Correlation coefficients resulting from LDA applied to normalized and standardized physical attributes of the Statistics-Based Regions in Set 1.

Physical Characteristic	LD_1	LD_2
Drainage Area	-0.409	0.190
Main Channel Slope	-0.001	0.191
Basin Slope	**-0.813**	-0.372
Basin Shape	-0.159	-0.292
Basin Elevation	**-0.549**	-0.306
% Impervious Cover	-0.066	0.535
% Forest Cover	**-0.561**	-0.120
Soil Drainage Index	**0.677**	0.075
Infiltration Index	0.315	-0.175
Variance Explained:	0.826	0.174

3.3.2 Physically-Based Regions

It is typically assumed that regions which are physically homogeneous are also hydrologically homogeneous. To validate this assumption and to confirm the results of the LDA above, Physically-Based Regions are evaluated using principal component and canonical correlation analyses applied to the normalized and standardized values of the physical variables for watersheds within Set 1. The PCA is used to identify the physical variables that explain most of the variability in the dataset, while the CCA will identify

the physical variables that most correlate to the flood statistics, and thus would be most appropriate for use in regional frequency analyses.

Principal Component Analysis

A principal component analysis is used to identify independent linear combinations of the physical variables which can be used to summarize the data set. Table 3.7 contains the principal component loadings corresponding to each normalized and standardized variable. For a given PC, each loading explains the importance of the associated variable. The higher the loading, the more impact the variable has on the principal component in question. The last two rows of Table 3.7 contain the eigenvalues associated with each PC and the cumulative variance explained by higher order PCs. According to the general rule of thumb that the PCs with eigenvalues greater than 1 should be retained, it is the first three PCs which are of interest here. These PCs explain nearly 68% of the variability in the dataset, with the first principal component (PC_1) alone explaining slightly more than PC_2 and PC_3 combined. Within PC_1, the most important variables are: basin slope, mean elevation, and soil drainage. PC_2 is largely driven by the drainage area and basin shape, and to a lesser degree, main-channel slope. PC_3 indicates that the land cover metrics (percent impervious surface and percent forest cover) are also influential. There is no doubt that all of these variables play some role in the hydrological process; however, these results alone do not indicate whether these variables are in fact indicative of hydrological homogeneity with respect to flood flows.

Table 3.7
Principal component loadings obtained for normalized and standardized physical variables at sites in Set 1.

Physical Characteristic	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7	PC_8	PC_9
Drainage Area	-0.167	**0.562**	0.000	-0.300	-0.256	0.432	0.540	-0.105	0.000
Main Channel Slope	0.284	**-0.451**	0.000	-0.445	0.000	-0.392	0.506	-0.174	0.263
Basin Slope	**0.497**	0.199	0.107	0.000	0.210	0.213	-0.254	0.155	0.722
Basin Shape	0.000	**0.607**	0.000	-0.221	0.241	-0.706	0.000	0.111	0.000
Basin Elevation	**0.445**	0.000	0.000	0.185	0.564	0.172	0.450	0.196	-0.414
% Impervious Cover	-0.105	0.000	**0.729**	0.503	0.000	-0.177	0.202	-0.303	0.193
% Forest Cover	0.321	0.000	**0.552**	-0.328	-0.444	0.000	-0.210	0.347	-0.344
Soil Drainage Index	**-0.482**	-0.205	0.127	0.000	0.123	0.000	0.225	0.761	0.255
Infiltration Index	-0.311	-0.168	0.349	-0.515	0.541	0.228	-0.216	-0.305	0.000
Eigenvalues:	1.769	1.325	1.132	0.934	0.802	0.719	0.679	0.489	0.319
Cumulative Variance:	0.348	0.543	0.685	0.782	0.853	0.919	0.962	0.988	1.000

Canonical Correlation Analysis

A Canonical Correlation Analysis is used to identify the physical variables which are most correlated with flood statistics, and thus would be useful in delineating hydrologically homogeneous regions. Herein, CCA was used to relate the nine normalized and standardized physical attributes to the first three L-moment ratios (L-CV, L-Skewness, and L-Kurtosis). Two canonical variates were created: U is the physical canonical variate consisting of the nine physical variables, and V is the hydrological

canonical variate consisting of the three L-moment ratios. Table 3.8 and Table 3.9 report the canonical correlations, and the coefficients of the physical and hydrological canonical variates as they relate to the actual physical and hydrologic variables, respectively. The maximum correlation coefficient that could be attained for the first pair of canonical variables (U_1,V_1) is 0.57. As indicated by the coefficients in Table 3.9, the hydrological variate V_1 is heavily impacted by the L-CV. Because the L-CV is the key component of hydrological homogeneity as defined by Hosking and Wallis (1997), the coefficients of U_1 in Table 3.8 indicate which physical variables are most relevant in characterizing the flood regime. These variables, by order of importance, are: basin slope, soil drainage, mean basin elevation, and percent forest cover.

Table 3.8
Canonical correlations and coefficients of physical variates defined for Set 1.

Physical Characteristic	**U_1**	**U_2**	**U_3**
Drainage Area	0.090	-0.564	-0.433
Main Channel Slope	0.210	0.331	0.365
Basin Slope	**0.953**	0.039	-0.044
Basin Shape	0.165	-0.283	-0.150
Basin Elevation	**0.733**	-0.147	0.459
% Impervious Cover	-0.090	-0.526	0.428
% Forest Cover	**0.625**	0.153	-0.023
Soil Drainage Index	**-0.816**	-0.391	-0.164
Infiltration Index	-0.369	-0.187	0.053
Correlation(U,V):	0.571	0.437	0.271

Table 3.9
Canonical correlations and coefficients of the hydrological variates defined for Set 1.

Hydrologic Variable	V_1	V_2	V_3
L-CV	**-0.888**	0.171	0.427
L-Skewness	-0.314	0.696	0.645
L-Kurtosis	-0.234	0.962	0.142
Correlation(U,V):	0.571	0.437	0.271

3.3.3 Selected Attributes for Set 1

The results of the LDA, PCA and CCA performed on various combinations of flood statistics and physical variables all concur in that the basin slope, soil drainage, and basin elevation are the most indicative of hydrological homogeneity with respect to flood flows in the western portion of the Southeastern U.S. (Set 1). The percent forest cover is also an important indicator of the flood regime, although to a lesser degree than the other three variables. The weights on percent forest cover in U_1 are on the order of 15% to 40% less than the weights on basin slope, soil drainage, and elevation. Therefore, percent forest cover will not be employed as an attribute in the subsequent steps to delineate homogenous regions presented in the following section.

3.4 Region Delineation in Set 1 Using Standardized Procedure

An application of the standardized procedure for delineation of hydrologically homogeneous regions is presented herein using the sites contained in Set 1. Based on the results presented in Section 3.3, mean basin elevation (E), mean basin slope (S_B), and the

soil drainage index (SI) are the most indicative of the flood regime in this study area. Therefore, the normalized and standardized values of those three physical variables are used as attributes in the Ward clustering technique to delineate regions. The resulting dendrogram is presented in Figure 3.4. Seven regions should be created to ensure that each region contains a minimum of 10 sites, while simultaneously minimizing the error sum of squares (or height on y-axis in Figure 3.4). Figure 3.5 shows the spatial distribution of these seven regions. Overall, the spatial continuity of each region is sufficient for visual classification of ungauged sites within the study area.

For each region, Table 3.10 reports the number of sites therein (N), the heterogeneity measure (H), and the total record length available ($\sum_{all\ i} n_i$). These metrics are presented for the original regions resulting from the Ward clustering technique, as well as for regions modified by removing sites identified as hydrologically discordant (HD), and for regions modified by removing only the sites which are both hydrological and physically discordant (HPD). The discordant sites are also noted in Figure 3.5, wherein circles denote HD sites and squares denote HPD sites; the color of the symbol corresponds to the original region within which the site in question was classified.

Five of the original regions (1, 3, 4, 5, and 6) are sufficiently homogeneous with $H < 4$, and thus, subsequent procedures to identify and remove discordant sites therein are unnecessary. In fact, the heterogeneity of Region 6 increases from 3.70 to 4.83 following the removal of the HPD sites; this increase in heterogeneity suggests that removal of discordant sites is unnecessary when $H < 4$ in the original region delineation. However, efforts should be made to improve the homogeneity of Regions 2 and 7, if possible.

Unfortunately, Region 7 does not contain any sites which are either HD or HPD, and thus the region should be used with caution in subsequent derivations of quantile estimators for ungauged sites classified therein. Region 2 contains four HD sites, the removal of which reduces the H-statistic from 6.70 to 2.85, thereby creating a region which is sufficiently homogeneous. However, removal of sites which are only hydrologically discordant is ill-advised in the interest of applications with ungauged sites. Only one of the four sites identified as hydrologically discordant in Region 2 is also physically discordant. Removal of this HPD site results in a more modest reduction in the H-statistic from 6.70 to 5.06, and thus the region remains reasonably heterogeneous. In all cases, the total record length available is greater than 500 years, and therefore, a sufficient amount of gauged data is available for estimation of the 100-year flood.

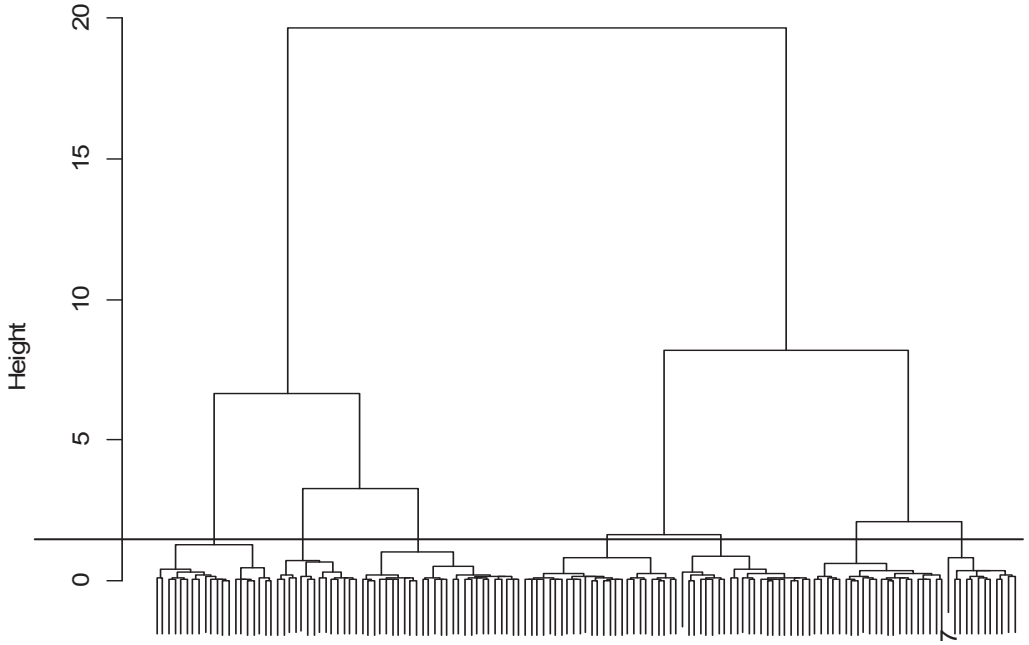


Figure 3.4: Dendrogram for Wards clustering applied to E, S_B, and SI in Set 1.

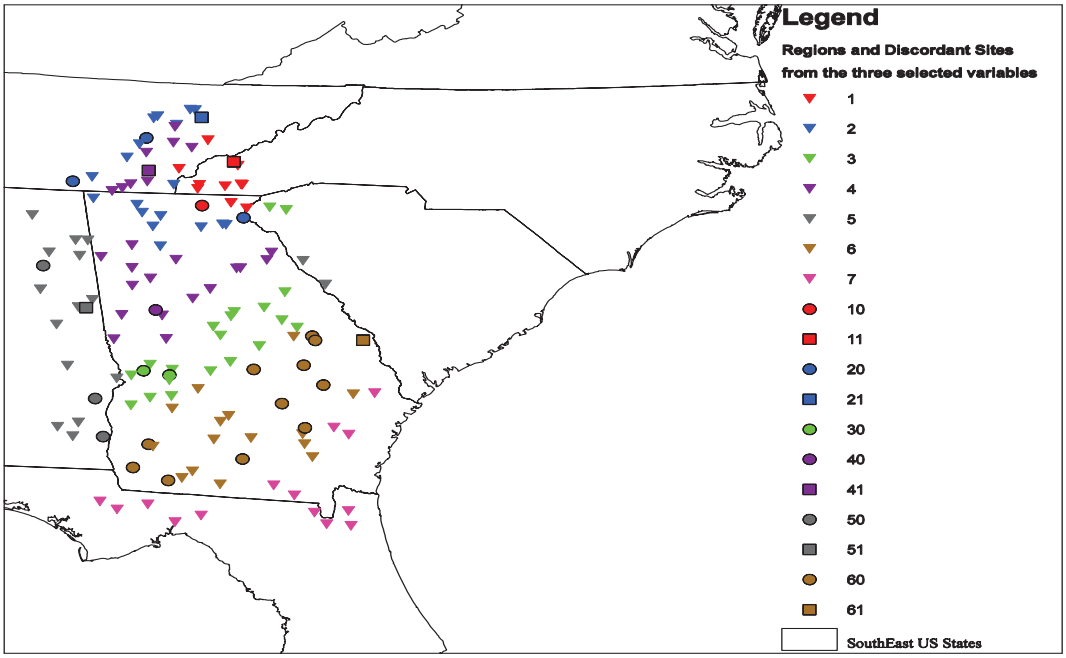


Figure 3.5: Seven regions delineated for Set 1 using Wards clustering applied to normalized and standard values of E, S_B, and SI. Circles represent HD sites, and squares represent HPD sites.

Table 3.10

Size and homogeneity of regions delineated for Set 1 using Wards clustering applied to normalized and standardized values of E, S_B, and SI.

	Original Regions			**Regions without HD sites**			**Regions without HPD sites**		
	N	H	$\sum_{all\ i} n_i$	N	H	$\sum_{all\ i} n_i$	N	H	$\sum_{all\ i} n_i$
Region 1	12	2.10	853	10	2.31	757	11	1.74	817
Region 2	22	6.70	1169	18	2.85	979	21	5.06	1131
Region 3	22	0.74	745	20	1.67	687	22	0.74	745
Region 4	26	3.09	1141	24	1.15	1054	25	2.26	1099
Region 5	20	2.55	960	16	2.36	824	20	1.93	935
Region 6	27	3.70	1056	15	4.03	556	26	4.03	984
Region 7	14	4.95	611	14	4.77	611	14	4.95	611

3.5 Accuracy of Quantile Estimators for Ungauged Basins in Set 1

The overall goal of regional flood frequency procedures is to develop flood quantile estimators for ungauged basins. As such, the accuracy of quantile estimators for ungauged basins classified within the regions delineated in Section 3.4 need to be evaluated to fully assess the efficiency of the proposed standardized procedure. To put the results in perspective, the performance of the quantile estimators derived using the standardized approach will be compared to that of estimators derived for regions delineated using methods commonly employed in practice. Overall, quantile estimators derived for five cases (or methods) will be considered. The cases differ primarily in terms of the similarity measures used to infer hydrologic homogeneity. Except for the use of alternative similarity measures employed in the Ward clustering technique, the proposed standardized procedure outlined in Section 3.1 is adhered to in most cases; however, in some cases additional multivariate statistical techniques are employed to further manipulate the physical data employed for clustering The following five cases are considered:

Case 1: Statistics-Based Regions are delineated using the at-site L-CVs as similarity measures in the Ward clustering technique (Saf 2009). Although flood statistics should not be used to delineate regions, this case will represent the best case scenario for hydrologic similarity. Results from this case will serve as the baseline to compare the following four physically-based cases against.

Case 2: Physically-Based Regions are delineated using the Ward clustering technique applied with all nine normalized and standardized physical variables

considered herein. This case represents the method most often employed in practice and has been recommended in many studies (e.g., Zrinji and Burn 1994; Burn 1997; Burn et al. 1997; Chiang et al. 2002a, 2002b; Rao and Srinivas 2006; Srinivas et al. 2008).

Case 3: Physically-Based Regions are delineated using the Ward clustering technique wherein principal component scores serve as attributes. The principal component scores are obtained from a prior principal component analysis applied to all nine normalized and standardized physical variables. This procedure is commonly applied in practice and takes advantage of the reductionist aspect of PCA where only the first few components are used in the regionalization (Alcazar and Palau 2010).

Case 4: Physically-Based Regions are delineated using the Ward clustering technique wherein the canonical scores related to the physical variate serve as attributes. The canonical scores are obtained by relating the L-CV, L-Skewness and L-Kurtosis to the nine normalized and standardized physical variables. Although applied infrequently in practice, this technique is still of interest (Tsakiris et al. 2011).

Case 5: Physically-Based Regions are delineated using the Ward clustering technique wherein the normalized and standardized values of mean basin elevation, mean basin slope, and the soil drainage index are employed as attributes. This corresponds to the standardized procedure for region delineation proposed herein.

The five cases are compared using the mean square error (MSE) of the quantile estimators for the 10-, 25-, 50- and 100-year events ($Q_{0.90}$, $Q_{0.96}$, $Q_{0.98}$, and $Q_{0.99}$, respectively). To assess the application of quantile estimators derived for ungauged sites

within a given region, the MSE is computed using a jackknife resampling procedure wherein, one at a time, each site is temporarily considered to be ungauged and removed from the region. For a given region, the MSE for a given return period, or corresponding cumulative probability *p*, is computed as:

$$MSE = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{\hat{Q}_p^R(i)-\hat{Q}_p^S(i)}{\hat{Q}_p^S(i)}\right)^2 \quad (56)$$

wherein N is the total number of sites contained in the region, $\hat{Q}_p^R(i)$ is the p[th] quantile estimator for site i derived using the GEV/L-moment Index Flood procedure, and $\hat{Q}_p^S(i)$ is the p[th] quantile estimator for site i derived using an at-site frequency analysis wherein the parameters of the GEV distribution are estimated as a function of L-moments. Normalization by $\hat{Q}_p^S(i)$ corrects for differences in scale across cases. (See for example, Chebana and Ouarda 2008.)

In addition, we propose a new metric, denoted *H**, for comparison of the heterogeneity across cases. As the use of different attributes in the clustering process is likely to yield a different delineation of regions, it is difficult to directly compare the H-statistics computed for the multiple regions delineated in each case. *H** represents the average heterogeneity value across all K regions delineated for a given case and is computed as

$$H^* = \frac{\sum_{k=1}^{K}(\sum_{all\ i} n_i)_k H_k}{\sum_{k=1}^{K}(\sum_{all\ i} n_i)_k} \quad (57)$$

where H_k is the heterogeneity measure of Region k, and $(\sum_{all\ i} n_i)_k$ is the total record length available in Region k. This metric will allow for a more direct comparison of the

ability of each case to delineate hydrologically homogeneous regions. Results for each case applied to the sites in Set 1 are discussed and compared in the following sections.

3.5.1 Results for Case 1

In this case, Statistics-Based Regions are delineated by employing the at-site L-CVs in Set 1 as similarity measures in the Ward clustering technique. The delineated regions are equivalent to those obtained in Section 3.3.1, wherein three regions were delineated to ensure that each contained a minimum of 10 sites (see Figure 3.2). The spatial representation of the three regions is shown in Figure 3.3. All of the delineated regions were sufficiently homogeneous with $H < 4$, and thus additional procedures to identify and remove hydrologically and physically discordant sites were not necessary (see Table 3.5). The efficiency of these regions for regionalization of flood data and applications at ungauged sites can be assessed using the MSEs of various flood quantiles illustrated in Figure 3.6. Additional results are tabulated in Appendix C. The highest mean square error obtained is 3.7% corresponding to estimation of the 100-year event in Region 1; this is likely due to Region 1 having the highest value of the H-statistic. On average, smaller MSEs are obtained for more frequent floods.

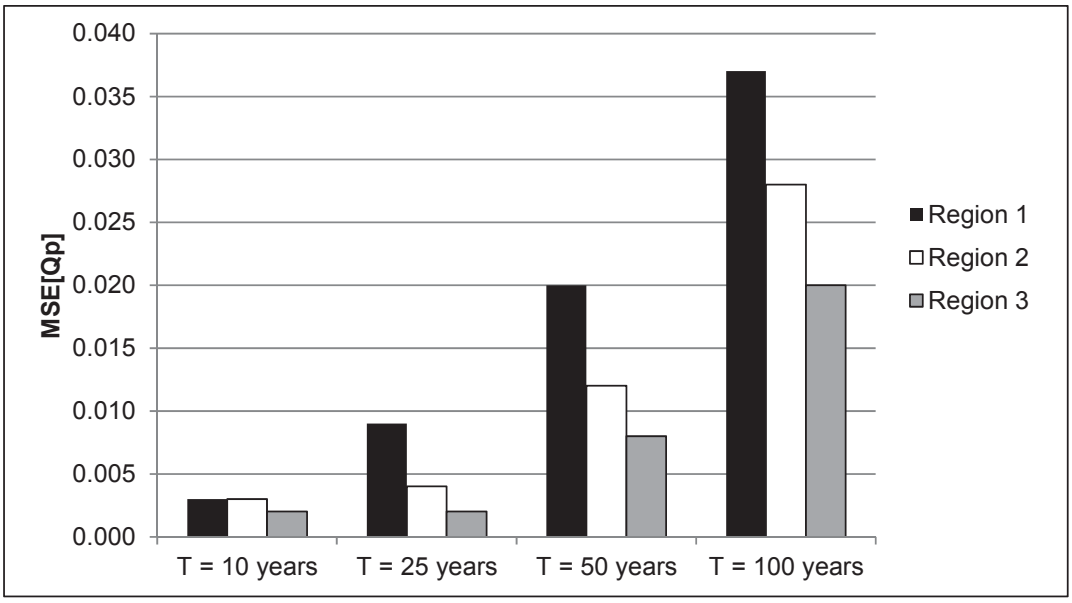


Figure 3.6: MSE of flood quantiles obtained for regions delineated in Case 1.

3.5.2 Results for Case 2

In this case, Physically-Based Regions are delineated by employing all nine normalized and standardized physical variables for sites in Set 1 as similarity measures in the Ward clustering technique. The five regions illustrated in Figure 3.7 were created to ensure that each contains a minimum of 10 sites. The delineated regions show a significant level of spatial continuity indicating a strong geographical component.

For each region, Table 3.11 reports the number of sites therein (N), the heterogeneity measure (H), and the total record length available ($\sum_{all\ i} n_i$). These metrics are presented for the original regions resulting from the Ward clustering technique, as well as for regions modified by removing HD sites, and for regions modified by removing HPD sites. The locations of these discordant sites are indicated in Figure 3.7. Regions 2, 3, and 4 are sufficiently homogeneous with H < 4; however, Regions 1 and 5

remain heterogeneous even after HPD sites are removed. Overall, the regions delineated using all available physical variables are more heterogeneous than regions delineated using only the mean basin elevation, mean basin slope, and soil drainage index (see Table 3.10). This suggests that the recommendation to cluster only on those variables which are most indicative of the flood regime is in fact appropriate.

Figure 3.8 illustrates the mean square errors for various quantile estimators at ungauged sites derived using the five regions delineated in Case 2. Results are for either the original regions resulting directly from the Ward clustering scheme in the instance that $H < 4$, or those which have been modified by removing the HPD sites in the instance that $H \geq 4$. Additional results are tabulated in Appendix C. The highest MSE obtained is 13.5% corresponding to estimation of the 100-year event in Region 5. As was observed in Case 1, on average, smaller MSEs are obtained for more frequent floods.

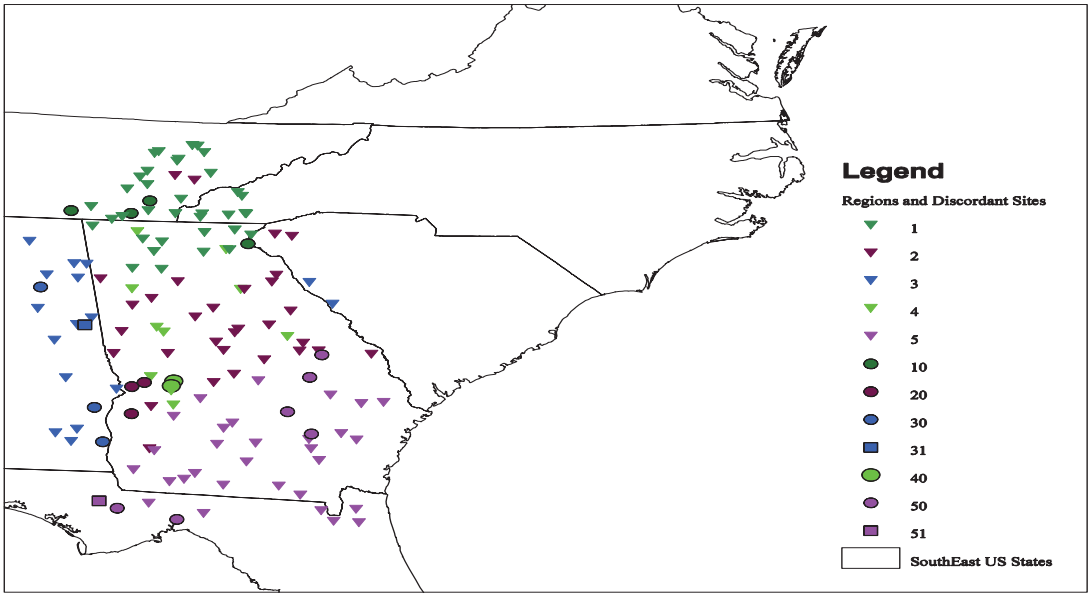


Figure 3.7: Five regions delineated for Set 1 using Wards clustering applied to all nine normalized and standardized physical variables. Circles represent HD sites, and squares represent HPD sites.

Table 3.11
Size and homogeneity of regions delineated for Set 1 using Wards clustering applied to all nine normalized and standardized physical variables.

	Original Regions			**Regions without HD sites**			**Regions without HPD sites**		
	N	H	$\sum_{all\ i} n_i$	N	H	$\sum_{all\ i} n_i$	N	H	$\sum_{all\ i} n_i$
Region 1	40	8.44	2238	36	5.87	2061	40	8.44	2238
Region 2	34	1.17	1347	31	-0.17	1254	34	1.17	1347
Region 3	20	2.62	960	16	2.40	824	19	1.93	935
Region 4	12	2.48	524	10	3.31	463	12	2.48	524
Region 5	37	6.09	1466	30	5.36	1141	36	5.77	1393

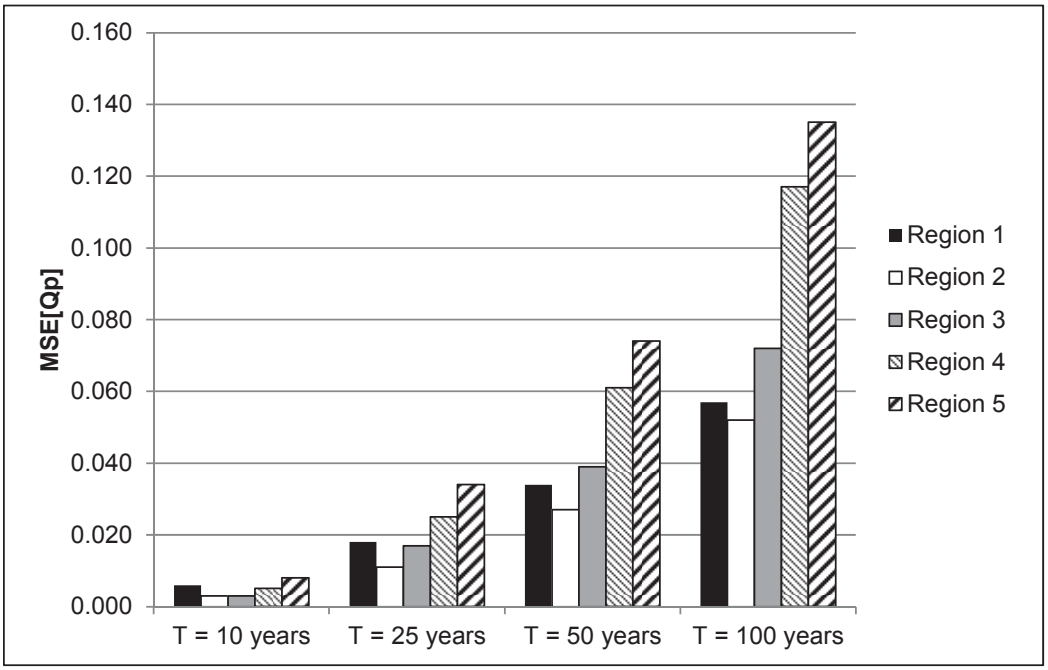


Figure 3.8: MSE of flood quantiles obtained for regions delineated in Case 2.

3.5.3 Results for Case 3

Instead of using the physical variables directly in the Ward clustering technique as in Case 2, herein the variables are first transformed into linear combinations (PCs) using a principal component analysis applied to all nine normalized and standardized physical variables for sites in Set 1. The resulting PCs are presented in Table 3.7. As the first three PCs have eigenvalues greater than one, the corresponding principal component scores computed for each site are used as similarity measures in the Ward clustering technique. The five regions illustrated in Figure 3.9 were created to ensure that each contains a minimum of 10 sites. These regions are not as contiguous as those delineated in Case 2, and are certainly not sufficient for visual classification of ungauged sites.

For each region, Table 3.12 reports the number of sites therein (N), the heterogeneity measure (H), and the total record length available ($\sum_{all\ i} n_i$). These metrics are presented for the original regions resulting from the Ward clustering technique, as well as for regions modified by removing HD sites, and for regions modified by removing HPD sites. The locations of these discordant sites are indicated in Figure 3.9. Overall, the regions obtained are less homogeneous than those obtained in Case 2.

Figure 3.10 reports the mean square errors for various quantile estimators at ungauged sites derived using the five regions delineated in Case 3. Results presented are for the original regions delineated as no HPD sites were identified. Additional results are tabulated in Appendix C. The highest MSE obtained is 15.6% corresponding to estimation of the 100-year event in Region 4. Overall, the MSEs for Case 3 are larger

than those observed for both Cases 1 and 2. Thus, Case 3 does not seem to be an acceptable approach for the regionalization of flood data.

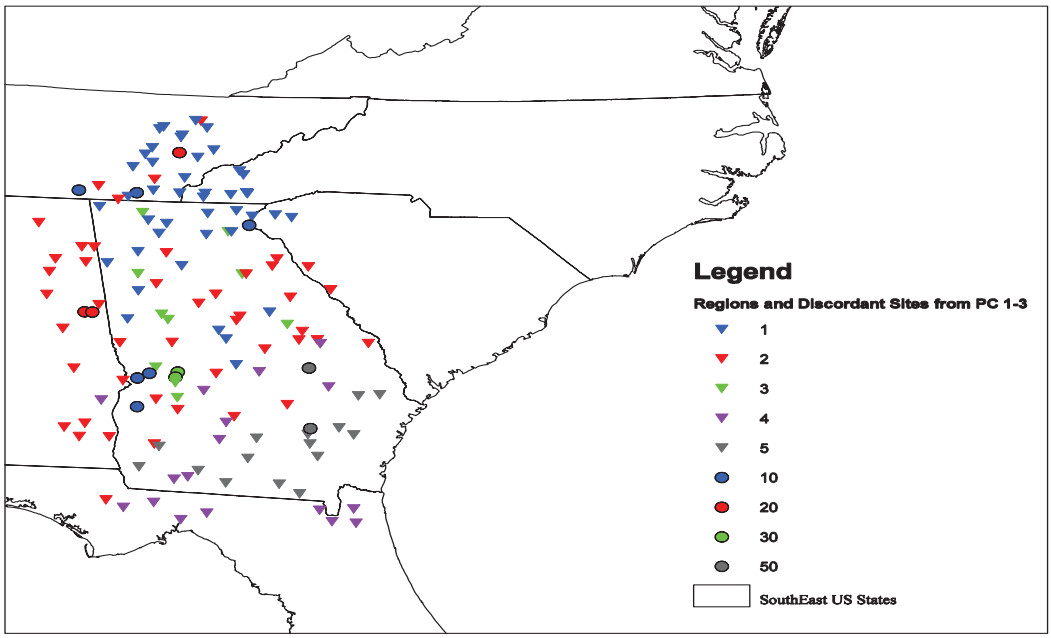


Figure 3.9: Five regions delineated for Set 1 using Wards clustering applied to PC_1, PC_2 and PC_3 created using the nine normalized and standardized physical variables. Circles represent HD sites, and squares represent HPD sites.

Table 3.12

Size and homogeneity of regions delineated for Set 1 using Wards clustering applied to PC_1, PC_2 and PC_3 created using the nine normalized and standardized physical variables.

	Original Regions			**Regions without HD sites**			**Regions without HPD sites**		
	N	H	$\sum_{all\ i} n_i$	N	H	$\sum_{all\ i} n_i$	N	H	$\sum_{all\ i} n_i$
Region 1	49	9.24	2331	43	6.65	2103	49	9.24	2331
Region 2	48	4.26	2335	45	3.74	2249	48	4.26	2335
Region 3	12	2.33	524	10	3.14	463	12	2.33	524
Region 4	17	3.00	671	17	3.00	671	17	3.00	671
Region 5	17	3.99	674	15	3.62	612	17	3.99	674

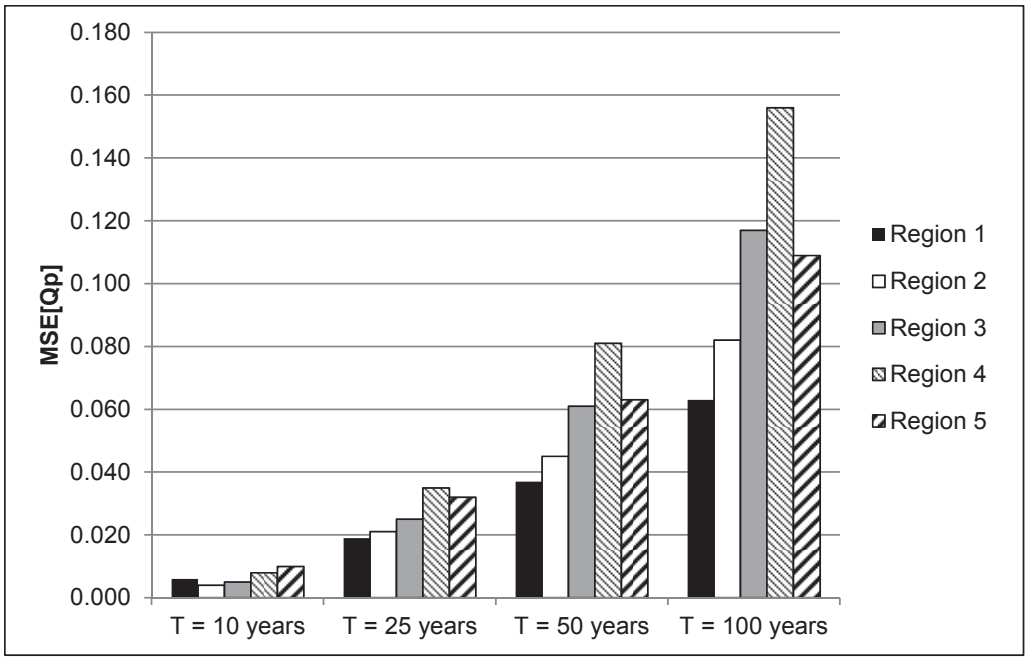


Figure 3.10: MSE of flood quantiles obtained for regions delineated in Case 3.

3.5.4 Results for Case 4

As in Case 3, the physical variables are transformed into linear combinations for use in the Ward clustering technique. In this case, a canonical correlation analysis is applied to two sets of variables for sites in Set 1: one contains all nine normalized and standardized physical variables, and the other contains the L-CV, L-Skewness, and L-Kurtosis. The resulting canonical variables are presented in Table 3.8 and Table 3.9. The canonical correlations indicate that the physical variate U_1 is the most relevant to the flood regime. Therefore, the coefficients for U_1 were used to compute canonical scores for each site, which in turn were used as similarity measures in the Ward clustering technique. The five regions illustrated in Figure 3.11 were created to ensure that each

contains a minimum of 10 sites. As was observed in Case 2, the spatial continuity of the delineated regions is sufficient to allow for visual classification of ungauged sites.

Table 3.13 reports the number of sites (N) contained in each region, the corresponding value of the H-statistic, and the total record length available ($\sum_{all\ i} n_i$). These metrics are presented for the original regions resulting from the Ward clustering technique, as well as for regions modified by removing HD sites, and for regions modified by removing HPD sites. The locations of these discordant sites are indicated in Figure 3.11. Overall, the regions obtained are more homogeneous than those obtained in Cases 2 and 3, but more heterogeneous than the regions obtained in Case 1. This result was expected as Case 4 is a more focused clustering approach than Case 2, because the canonical scores place more weight on the physical variables which are most indicative of hydrological homogeneity.

Figure 3.12 illustrates the mean square errors for various quantile estimators at ungauged sites derived using the five regions delineated in Case 4. Results are for either the original regions resulting directly from the Ward clustering scheme in the instance that $H < 4$, or those which have been modified by removing the HPD sites in the instance that $H \geq 4$. Additional results are tabulated in Appendix C. The highest MSE obtained is 12.2% corresponding to estimation of the 100-year event in Region 5. Overall, the MSEs obtained for Case 4 are lower than those observed for both Cases 2 and 3. Thus, of the physically-based techniques currently employed in practice, the use of canonical scores as attributes in the clustering algorithm yields regions which provide the most accurate quantile estimators for ungauged sites. Unfortunately, this method is used the least.

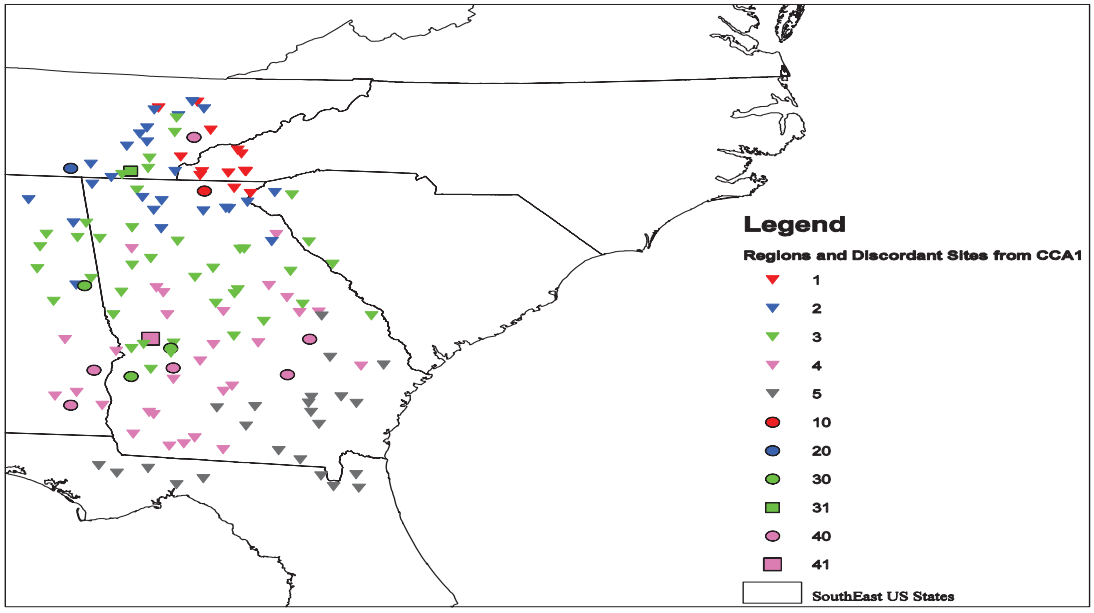


Figure 3.11: Five regions delineated for Set 1 using Wards clustering applied to the first physical canonical variate. Circles represent HD sites and squares represent HPD sites.

Table 3.13

Size and homogeneity of regions delineated for Set 1 using Wards clustering applied to the first physical canonical variate.

	Original Regions			**Regions without HD sites**			**Regions without HPD sites**		
	N	H	$\sum_{all\ i} n_i$	N	H	$\sum_{all\ i} n_i$	N	H	$\sum_{all\ i} n_i$
Region 1	14	3.75	978	13	4.03	918	14	3.75	978
Region 2	26	6.07	1419	24	5.13	1334	26	6.07	1419
Region 3	44	3.99	1759	40	2.66	1646	43	3.99	1728
Region 4	36	2.3	1427	29	2.56	1175	35	2.26	1379
Region 5	23	4.68	952	23	4.68	952	23	4.68	952

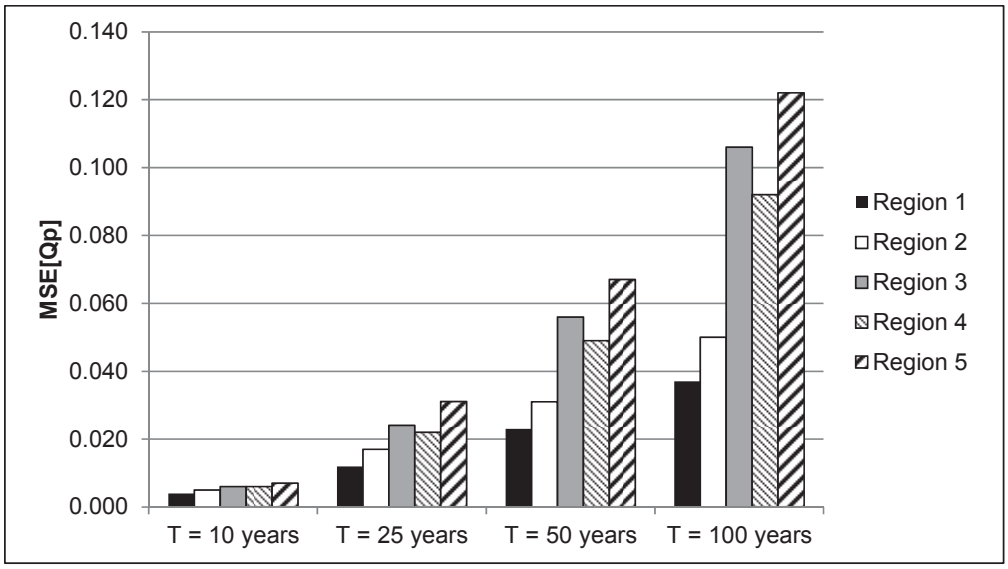


Figure 3.12: MSE of flood quantiles obtained for regions delineated in Case 4.

3.5.5 Results for Case 5

This case corresponds to the proposed standardized method for delineation of hydrologically homogeneous regions. In Section 3.3, it was concluded that the mean basin elevation (E), mean basin slope (S_B), and soil drainage index (SI) are most indicative of the flood regime in Set 1. The regions delineated using those three physical variables and the corresponding metrics related to regional heterogeneity are presented in Section 3.4 (see Figure 3.5 and Table 3.10). The efficiency of these regions for quantile estimation at ungauged sites is assessed herein using the MSEs of various flood quantiles illustrated in Figure 3.13. Results are for either the original regions resulting directly from the Ward clustering scheme in the instance that $H < 4$, or those which have been modified by removing the HPD sites in the instance that $H \geq 4$. Additional results are tabulated in Appendix C. The highest MSE obtained is 13.2% corresponding to estimation of the 100-year event in Region 6. On average, the MSEs obtained for Case 5

are slightly less than those observed for Case 4; therefore, the newly proposed standardized method for region delineation is the most competitive physically-based method for quantile estimation at ungauged sites. As with Case 4, Case 5 is a more focused clustering approach than Case 2; however, Case 5 employs three attributes (E, S_B, and SI) in the Ward clustering technique, whereas Case 4 employs only one attribute (U_1). A more detailed comparison of the five cases is provided in the next section.

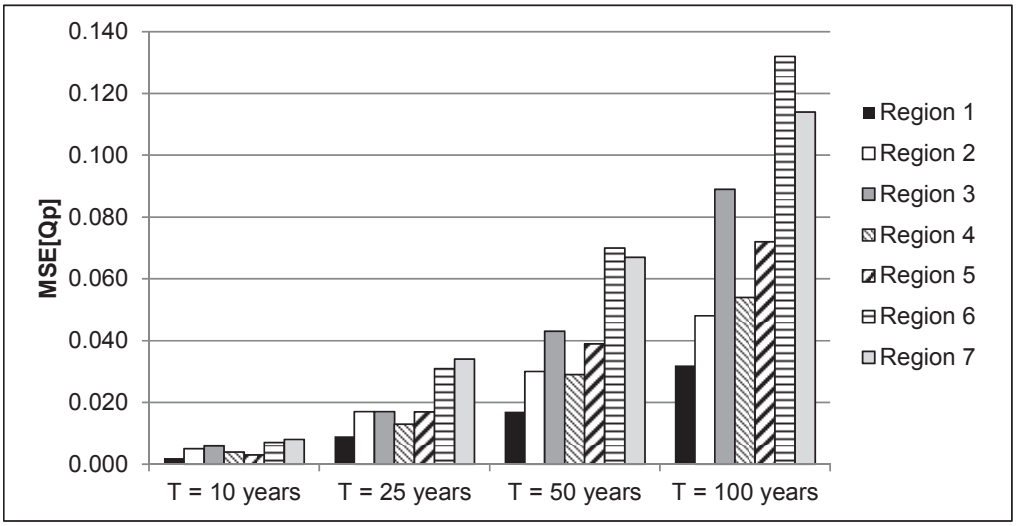


Figure 3.13: MSE of flood quantiles obtained for regions delineated in Case 5.

3.5.6 Comparison of Cases

Results presented above indicate that Case 5, corresponding to the standardized approach for region delineation presented herein, yields the most accurate quantile estimators for use in ungauged basins relative to other physically-based approaches typically employed in practice. Tables presented in this section illustrate this finding more clearly. Table 3.14 reports the average heterogeneity level (H*) for each case considered in regards to the original regions defined by the Ward clustering technique, as

well as regions from which all HD sites are removed, and regions from which only HPD sites are removed. For all three scenarios, Case 5 yields the most homogenous regions relative to the other physically-based procedures. Although the H-statistics for Case 5 are not as competitive as those obtained in Case 1, all H* values obtained for Case 5 are less than 4 indicating that the regions are sufficiently homogeneous for the development of quantile estimators at ungauged sites. In addition, it should be noted that the H-statistics for Case 1 are exceptionally small due to the use of L-CVs for both region delineation and the subsequent evaluation of regional homogeneity.

Figure 3.14 illustrates the average MSEs computed for various flood quantiles obtained for regions delineated in Cases 1 through 5 and subsequently modified to remove all HPD sites as appropriate (i.e., if $H \geq 4$). The magnitudes of the MSEs are satisfactory in comparison with results from other regionalization studies such as Malekinezhad et al. (2011) and Shu and Ouarda (2009). Overall, Case 5 outperforms the other physically-based methods considered. And, in comparing the MSEs with the H* values in Table 3.14, it is evident that improved accuracy of quantile estimators for ungauged sites is a direct result of the increased hydrological homogeneity of delineated regions. Thus, use of the physical variables most indicative of the flood regime as attributes in the delineation process results in greater hydrologic homogeneity of delineated regions, and more efficient quantile estimators for use at ungauged sites are obtained.

Table 3.14
Average heterogeneity (H*) obtained for regions delineated in Cases 1-5.

	Case 1 (L-CV)	**Case 2** (All Vars)	**Case3** (PC 1-3)	**Case 4** (U_1)	**Case 5** (E, S_B, SI)
Original Regions	-1.15	5.08	5.54	6.81	3.53
Regions without HD sites	-1.73	3.75	4.60	3.72	2.56
Regions without HPD sites	-1.03	4.91	5.72	4.14	3.02

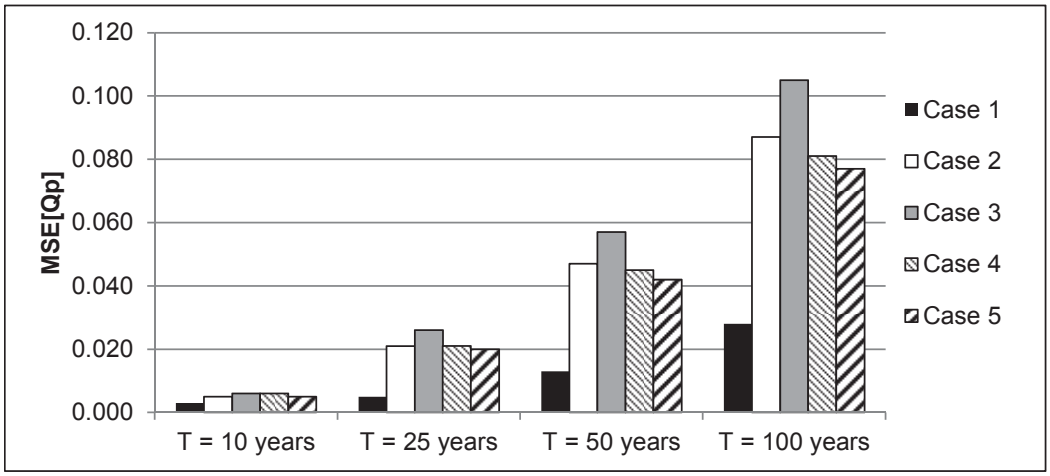


Figure 3.14: Average MSE of flood quantiles obtained for regions delineated in Cases 1 through 5. Results are for regions appropriately modified to remove HPD sites.

3.6 Validation of Standardized Procedure in Alternate Study Areas

The analyses above revealed that three physical variables (mean basin slope, mean basin elevation, and soil drainage) are sufficient to delineate hydrologically homogeneous regions in the western portion of the Southeastern U.S. (Set 1). A primary goal of this dissertation is to demonstrate that those same variables could be used to infer the flood regime in alternate study areas. The possible extrapolation of these results from Set 1 to Set 2 of the Southeastern U.S. is investigated herein. However, as noted in

Section 3.2, with the exception of drainage area, the physical characteristics of the watersheds in Sets 1 and 2 are statistically different on average; this may negatively impact the extrapolation of results.

To investigate whether the findings for Set 1 can be generalized to alternate study areas, normalized and standardized values of the mean basin slope, mean basin elevation, and soil drainage index were used as attributes in the Ward clustering technique to delineate regions in Set 2. With the exception of identifying the physical variables most indicative of the flood regime, regions are delineated using the standardized method proposed herein. Therefore, the six regions illustrated in Figure 3.15 were created to ensure that each contains a minimum of 10 sites, while simultaneously minimizing the error sum of squares. The spatial continuity of the delineated regions should be adequate for visual classification of ungauged sites.

Table 3.15 reports the number of sites (N) contained in each region, the corresponding value of the H-statistic, and the total record length available ($\sum_{all\ i} n_i$). These metrics are presented for the original regions resulting from the Ward clustering technique, as well as for regions modified by removing all HD sites, and for regions modified by removing only HPD sites. The locations of these discordant sites are indicated in Figure 3.15. With the exception of Region 1, the heterogeneity measures for the original regions are all less than 4 indicating that the regions are adequate for regionalization of flood data. In addition, Region 1 is at its core quite homogeneous, as removing five HD sites reduces the H-statistic to -0.4 and a sufficient number of years of record remains. However, it is recommended that these sites be retained for use in

subsequent steps to derive quantile estimators for ungauged sites as they are not physically discordant.

The jackknife resampling procedure discussed in Section 3.5 was also applied to the regions delineated for Set 2 in order to evaluate their adequacy for the development of quantile estimators for ungauged sites. The resulting MSEs for various quantiles are illustrated in Figure 3.16 for either the original regions resulting directly from the Ward clustering scheme in the instance that $H < 4$, or those which have been modified by removing the HPD sites in the instance that $H \geq 4$. Additional results are tabulated in Appendix C. Overall, the highest MSE obtained was 13.6%, corresponding to estimation of the 100-year event in Region 6. Both the MSEs and H* values obtained for regions delineated in Set 2 using basin slope, elevation, and soil drainage as similarity measures are competitive with the values obtained for regions delineated in Set 1 using the same attributes. As the physical characteristics of Sets 1 and 2 are statistically different on average, these results indicate that the three variables identified for Set 1 are sufficient to characterize the flood regime in alternate study areas.

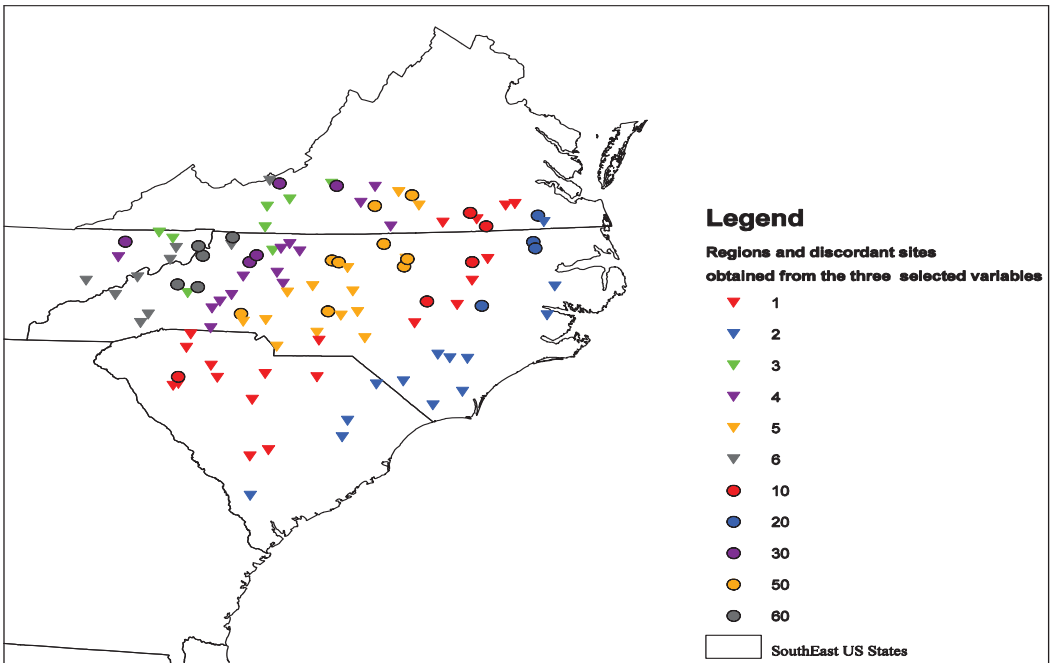


Figure 3.15: Five regions delineated for Set 2 using Wards clustering applied to normalized and standardized values of E, S_B, and SI. Circles represent HD sites and squares represent HPD sites.

Table 3.15

Size and homogeneity of regions delineated for Set 2 using Wards clustering applied to normalized and standardized values of E, S_B, and SI.

	Original Regions			**Regions without HD sites**			**Regions without HPD sites**		
	N	H	$\sum_{all\ i} n_i$	N	H	$\sum_{all\ i} n_i$	N	H	$\sum_{all\ i} n_i$
Region 1	25	4.52	1398	20	-0.45	1167	25	4.52	1398
Region 2	17	1.70	854	13	2.97	660	17	1.70	854
Region 3	13	2.29	845	8	0.42	533	13	2.29	845
Region 4	14	2.09	824	14	2.09	824	14	2.09	824
Region 5	23	3.30	1189	14	2.7	765	23	3.30	1189
Region 6	14	3.95	856	9	3.51	567	14	3.95	856
H*	3.1			1.6			3.1		

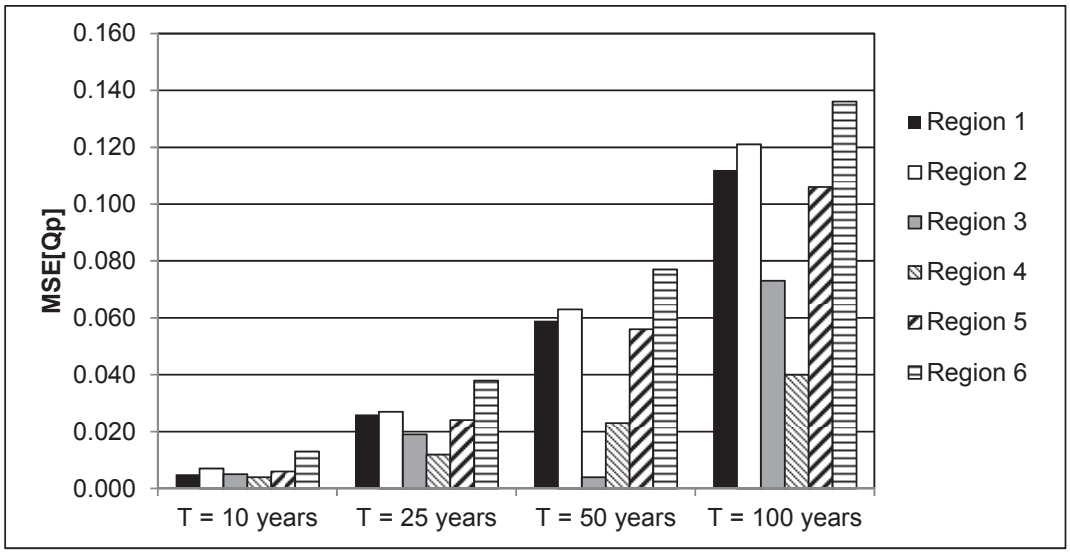


Figure 3.16: MSE of flood quantiles obtained for regions delineated in Set 2 as a function of E, SB, and SI.

3.7 Conclusion

This chapter presented a standardized procedure for the delineation of hydrologically homogeneous regions. A key aspect of this procedure is the identification of an appropriate set of physically based measures of hydrological similarity for use in region delineation. Overall, the results presented herein demonstrate that mean basin elevation, mean basin slope, and soil drainage are appropriate for regionalization of flood data throughout the Southeastern United States. Overall, use of these physical characteristics as attributes in Wards clustering technique yields regions which are more homogeneous than those delineated using physically-based procedures typically employed in practice. In addition, results of a jackknife resampling procedure presented herein demonstrate that increasing the homogeneity of delineated regions results in the ability to develop more accurate quantile estimators for use in ungauged basins. Further,

these results can be extrapolated to study areas with statistically different physical characteristics. These findings will be used in Chapter 5 to develop quantile estimators for ungauged basins in Haiti.

Chapter 4 : Delineation of Hydrologically Homogeneous Regions Using Spatially Distributed Data

With increased computing capabilities and the availability of data, spatially distributed models are becoming more prevalent (Beven and Kirkby 1979; Abbott et al. 1986; Boyle et al. 2001; Duffy 2004; Panday and Huyakorn 2004; Reed et al. 2007). As basins are composed of a wide range of landscape properties, use of spatially distributed parameters provides a better representation of the watershed, which in turn, is expected to represent hydrological processes more accurately. For this reason, distributed methods are often used in hydrologic modeling at the watershed scale instead of lumped rainfall-runoff models (see for example, Ajami et al. 2004; Paudel et al. 2009, and citations therein). On the contrary, most regionalization studies consider an individual watershed to be a homogeneous unit whose properties can be represented by a single aggregated variable. In the latter case, it is necessary to infer the dominant processes governing hydrologic response at the watershed scale such that simple models can be developed to extrapolate results from gauged basins for improved prediction in ungauged basins (e.g., McDonnell et al. 2007; Tezlaff et al. 2008; MacKinnon and Tetzlaff 2009).

The goal of this chapter is to determine whether or not physical values aggregated at the watershed scale are sufficient for regionalization of flood data. In particular, this chapter investigates whether regions delineated using spatially distributed variables yield more accurate quantile estimators for ungauged sites than those derived from regions delineated using aggregated values. In Chapter 3, it was observed that similarity in the flood regime throughout the Southeastern U.S. is primarily defined by three physical

characteristics: mean basin slope, mean basin elevation, and soil drainage. This conclusion was drawn based on results of multivariate statistical techniques applied to aggregated values of physical variables commonly employed as indicators of extreme hydrologic response. In this chapter, spatially distributed representations of the three key physical variables (E, S_B, and SI) are used to delineate hydrologically homogeneous regions; the accuracy of quantile estimators derived for use in ungauged basins therein are subsequently evaluated and compared to the results in Chapter 3. Overall, the only difference in the procedures employed herein versus that in Chapter 3 is due to the derivation of similarity measures, i.e., spatially distributed versus aggregated attributes.

Herein, unsupervised classification techniques are used to derive spatially distributed representations of the key physical variables (E, S_B, and SI). A similar type of analysis was conducted by Mazvimavi et al. (2006) wherein remotely sensed data was used to define additional parameter values related to land cover and the geological features of watersheds. Using redundancy analysis, they related the hydrologic behavior of watersheds to physical attributes describing the proportion of each watershed occupied by various land covers and rock types. However, their study only considered hydrologic response as measured by the mean annual runoff, the coefficient of variation of the annual runoff, percentiles of daily flow, and the average number of days per year without flow. The impact of spatially distributed variables on the regionalization of flood statistics has not previously been considered.

4.1 Unsupervised Classification

Two main parametric classification techniques are employed in remote sensing: supervised classification and unsupervised classification (Jensen 2005, p. 337). In general, each technique employs a classifier (similar to a clustering algorithm) to categorize pixels of a digital image into classes which are subsequently related to values of the physical variable of interest (e.g., land cover type). Supervised classification requires prior knowledge of the characteristics of the classes to be created. In the application herein, there is no a priori knowledge of the characteristics of each class; therefore, unsupervised classification is preferred. Unsupervised classification is a remote sensing technique that is widely used in pattern recognition for spatial data (Jensen 2005, p. 337). Studies by Duveiller et al. (2007) and Hutchinson et al. (2010) are examples of relevant applications for land classification. The two most commonly used clustering algorithms for unsupervised classification are the k-means method and the Iterative Self-Organizing Data (ISODATA) algorithm. In the k-means method, the number of classes is set a priori, whereas only the maximum number of classes to be created is set a priori in the ISODATA algorithm. For this reason, the ISODATA algorithm is generally preferred over the k-means method, and will be employed in the following analyses using functions in ERDAS IMAGINE (Erdas 2010). (Herein, a maximum of 14 classes are created for each physical variable as discussed in Section 4.2).

The ISODATA algorithm is an iterative procedure which ultimately yields classes (or groups of pixels) with similar spectral characteristics. Initial classes are formed by

randomly selecting a number of points along the range of possible values for the parameter in question. In each iteration, individual pixel values are compared to the mean of each class. Each pixel is assigned to the class for which the mean value is closest to that of the pixel. Once all pixels have been classified, the mean values of each class are recalculated. The new means are input into the next iteration, and pixels are reassigned to a different class as needed. This process is repeated until there is no significant change in pixel assignments or the maximum number of iterations is reached. In particular, the algorithm stops when the classes to which 95% of the pixels are assigned remain unchanged from one iteration to another, or a maximum of 6 iterations are performed. (These are default settings in ERDAS IMAGINE and will be used herein.)

In addition to the general process outlined above, the ISODATA algorithm includes the following criteria for the refinement of classes (Jensen, 2005):

1. If a given class does not contain at least 1% of the total number of pixels, then it is merged with the class with the closest mean.
2. If the minimum weighted distance between two classes is less than a predefined value, then the two classes are merged.
3. If the maximum standard deviation in a class exceeds a predefined value, then the class is subdivided to create two new classes for which the means are equivalent to that of the original class plus and minus one standard deviation.

The values of the critical/predefined thresholds used herein to determine when to merge or split classes are based on default values set within ERDAS.

The application of the ISODATA algorithm to obtain spatially distributed values of the elevation and basin slope for each site (and corresponding watershed) in Figure 3.1 is discussed below; spatially distributed values corresponding to soil drainage are readily available for the study area. Following convergence of the classification algorithm, a number of classes are created. In the application herein, additional analyses are needed to relate the classes to individual watersheds. For a given watershed, GIS is used to overlay the watershed boundary on a map representing the classes, and the percentage of the watershed falling within each class is calculated. The classes defined for each of the key physical variables (E, S_B, and SI) are subsequently used as similarity measures in the Ward clustering technique to delineate hydrologically homogeneous regions throughout the Southeastern U.S.

4.2 Derivation of Spatially Distributed Variables

A spatially distributed representation of the soil drainage within the Southeastern U.S. is presented in Figure 4.1. This data was obtained from the conterminous U.S. (CONUS) soil database which was designed for regional studies and contains information based on soil surveys from the State Geographic Soil (STATSGO) database with some minor modifications (http://www.soilinfo.psu.edu/). The spatial resolution of the STATSGO data is 6.25 km^2 (http://www.il.nrcs.usda.gov/technical/soils/statsgo_inf.html#statsgo5). As the minimum drainage area considered in the study is

approximately 9 km^2, the spatial resolution of the soil drainage data is considered appropriate for the spatial analysis performed herein.

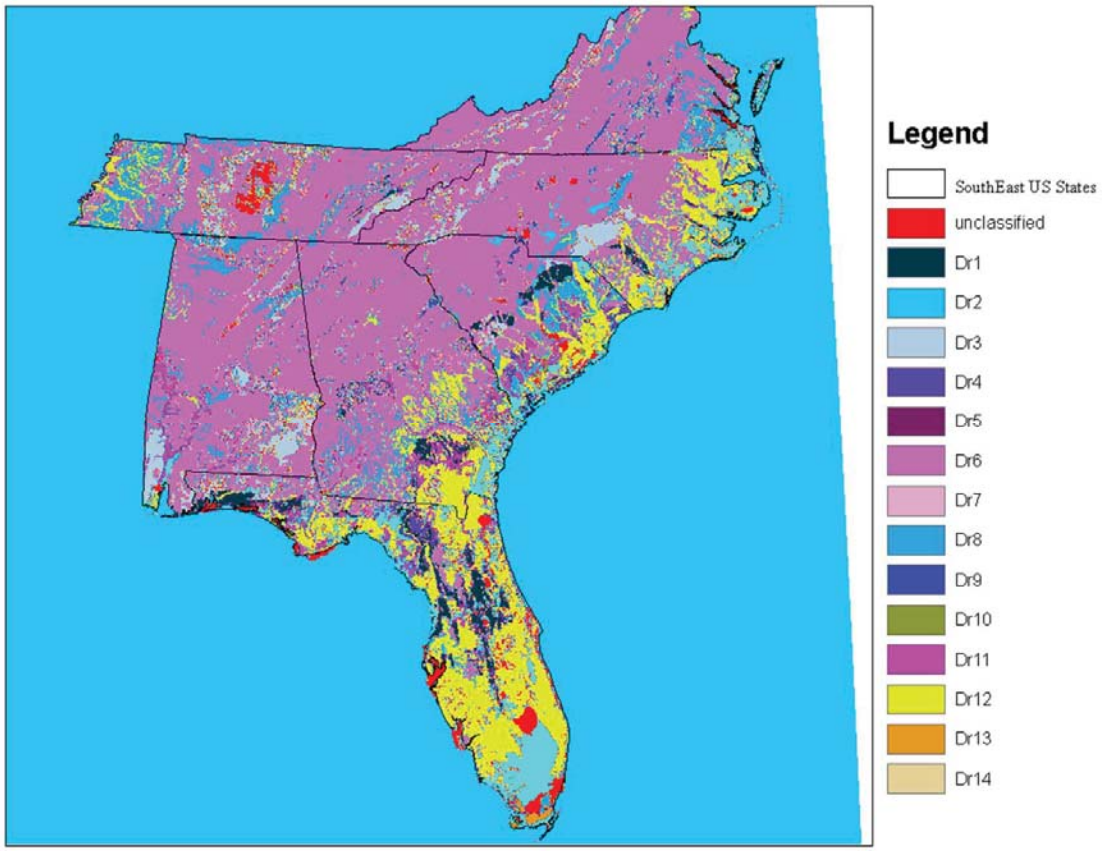


Figure 4.1: Spatial representation of the fourteen soil drainage classes in the Southeastern U.S. as obtained from STATSGO (http://www.soilinfo.psu.edu/).

The soil drainage is assembled into 14 classes corresponding to the seven major soil drainage classes as defined in Appendix A, as well as seven additional classes corresponding to mixed soil types. Herein, the proportion of each watershed contained in each of the 14 drainage classes (denoted Dr1 – Dr14) will serve as attributes in the Ward clustering technique. As such, an additional 13 attributes related to soil drainage will be used to delineate regions using spatially distributed variables as opposed to the single soil

drainage index used in Chapter 3. Physical descriptions of each drainage class and the corresponding soil drainage indices are provided in Table 4.1. The majority of the Southeastern U.S. is well drained and is contained in drainage class 1 (Dr1). Negligible portions of the dataset are unclassified (Dr0) or of mixed soil types.

Table 4.1
Characteristics of drainage classes for the Southeastern U.S.

Drainage Type	Description	Soil Drainage Index (SI)	Drainage Class	Proportion of Dataset
Unclassified	NA	NA	Dr0	3.30%
E	Excessively drained	1.0	Dr1	3.40%
W,E	Mixed	1.5	Dr2	0.43%
SE	Somewhat excessively drained	2.0	Dr3	4.80%
W,SE	Mixed	2.5	Dr4	0.50%
MW, SE	Mixed	2.7	Dr5	0.04%
W	Well drained	3.0	Dr6	55.5%
W,MW	Mixed	3.5	Dr7	0.30%
MW	Moderately well drained	4.0	Dr8	5.70%
MW,SP	Mixed	4.5	Dr9	0.60%
SP,MW	Mixed	4.5	Dr10	0.20%
SP	Somewhat poorly drained	5.0	Dr11	3.70%
P	Poorly drained	6.0	Dr12	15.2%
P, VP	Mixed	6.5	Dr13	0.20%
VP	Very poorly drained	7.0	Dr 14	6.10%

Spatially distributed values of elevation and basin slope for the Southeastern U.S. were derived using SRTM (Shuttle Radar Topography Mission) DEMs of 90m resolution (8100 m^2 per pixel) downloaded from the Consultative Group on International

Agricultural Research consortium for spatial information (CGIAR-CSI) website (http://srtm.csi.cgiar.org/). The decision to use DEMs of 90m resolution can be partially explained by the large spatial extent of the study area: a lower resolution is advantageous as it requires substantially less computational time to process the information. In addition, attempts to use 30m DEMs resulted in computational instabilities, and the 30m DEMs contain gaps within the study area which impaired subsequent analyses to characterize the spatial variability of the elevation, slope, and soil drainage at the watershed level (see below). The latter analyses require delineation of watershed boundaries, for which use of DEMs as coarse as 90m resolution is generally considered sufficient (Moglen and Hartman 2001; Pryde et al. 2007; Mendas 2010).

Mean elevations for each pixel are obtained directly from the DEMs. The mean slope of each pixel in degrees was calculated from the DEMs using the Spatial Analyst tool in ArcGIS. The spatial variability of the elevation and slope parameters was then characterized using unsupervised classification techniques. The unsupervised ISODATA algorithm was employed to create a maximum of fourteen classes for both elevation and slope. This maximum number of classes was chosen to ensure that no spatial variable was given a larger weight than the soil drainage. Using the ISODATA algorithm, the maximum of 14 classes were created for elevation (denoted E1 – E14); however, only eight classes were created for slope (denoted S1, S5, S6, S8, S9, S11, S12, and S14). In the latter case, classes were merged to ensure that each contained at least 1% of the total number of pixels. The spatial representations of the slope and elevation classes across the Southeastern U.S. are shown in Figure 4.2 and Figure 4.3, respectively; the physical

characteristics of each of the classes for slope and elevation are reported in Table 4.2 and Table 4.3, respectively. Most of the slopes in the Southeastern U.S. are less than 10 percent and are concentrated in classes S1 to S12; only class S14 contains slopes greater than 10 percent. With respect to elevation, nearly half of the study region is less than 13 m above mean sea level and is contained in class E1; the other half of the study region is roughly equally distributed among the remaining thirteen elevation classes.

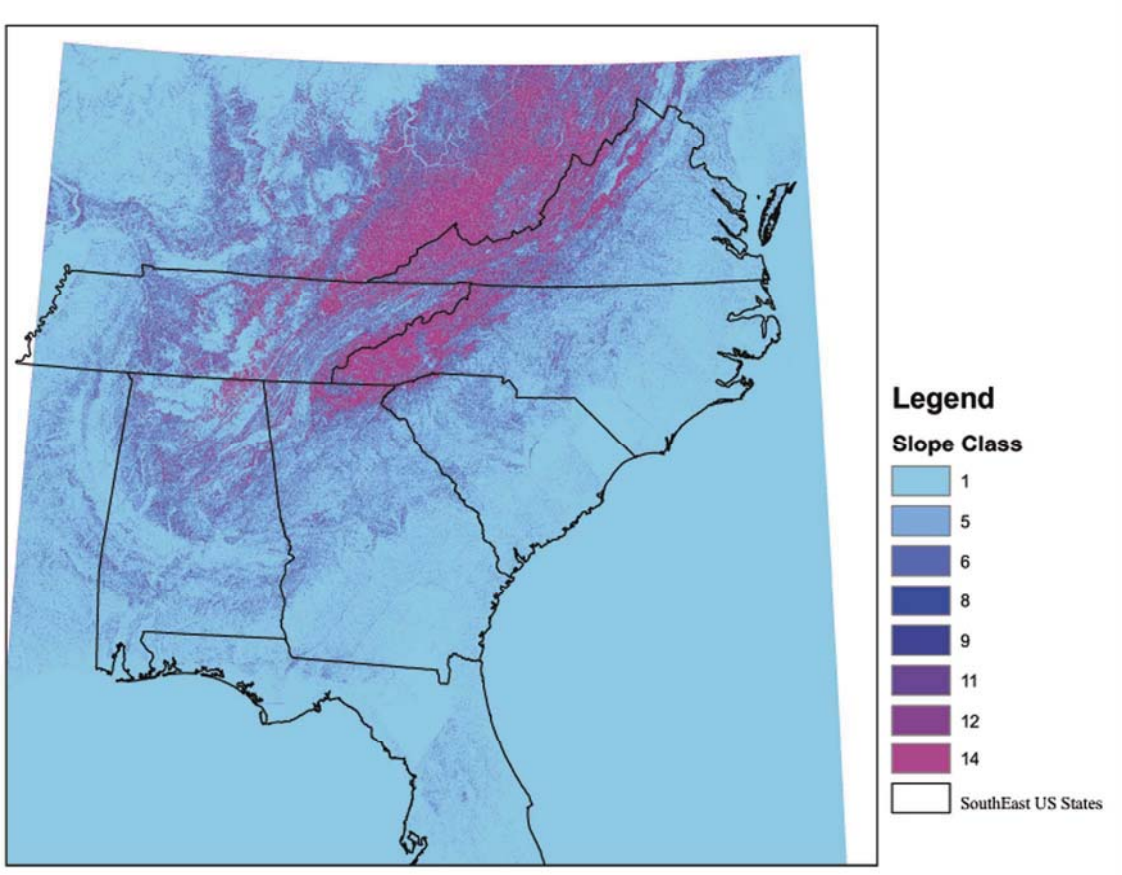


Figure 4.2: Spatial representation of eight slope classes for the Southeastern U.S. created using the ISODATA algorithm.

Table 4.2
Characteristics of slope classes for Southeastern U.S. created using ISODATA algorithm.

Percent Slope	Class	Proportion of Dataset
0 - 1	S1	77.0%
>1 - 2	S5	7.5%
>2 - 3	S6	4.1%
>3 - 4	S8	2.4%
>4 - 5	S9	1.6%
>5 - 6	S11	1.2%
>6 - 10	S12	2.4%
>10 - 77	S14	3.8%

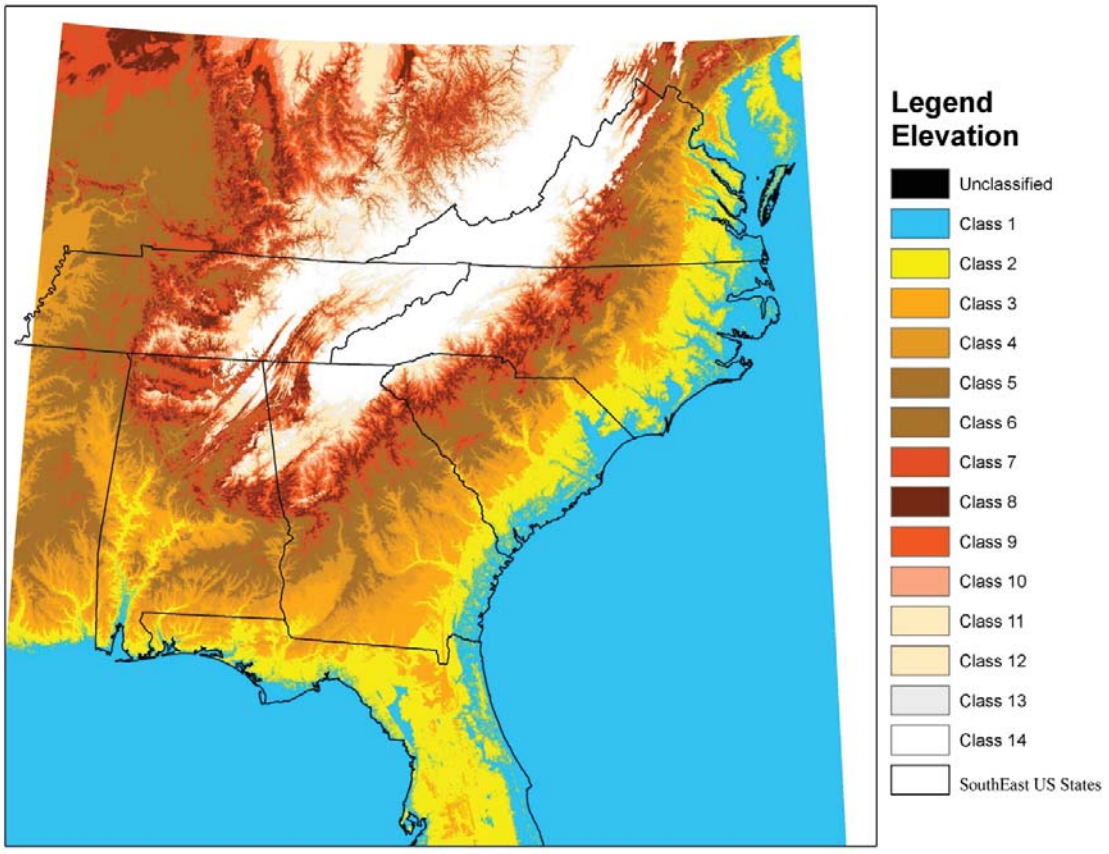


Figure 4.3: Spatial representation of fourteen elevation classes for the Southeastern U.S. created using the ISODATA algorithm.

Table 4.3
Characteristics of elevation classes for the Southeastern U.S. created using the ISODATA algorithm.

Elevation (meters)	Class	Proportion of Dataset
0 - 13	E1	47.2%
>13 - 42	E2	7.1%
>42 - 73	E3	4.7%
>73 - 105	E4	5.0%
>105 - 137	E5	5.5%
>137 - 168	E6	5.0%
>168 - 200	E7	4.1%
>200 - 231	E8	3.7%
>231 - 262	E9	3.4%
>262 - 293	E10	3.0%
>293 - 325	E11	2.2%
>325 - 356	E12	1.4%
>356 - 516	E13	3.1%
>516 - 2025	E14	4.4%

In order to characterize the spatial variability of the soil drainage, slope, and elevation at the watershed level, the watersheds corresponding to the 249 gauging stations in Figure 3.1 were delineated using the SRTM 90m DEM in ArcMap. Figure 4.4 shows the spatial extent of each watershed. Subsequently, the percentage of each watershed belonging to each previously defined drainage, slope, and elevation class (Dr1-Dr14, S1-S14, and E1-E14) was ascertained by overlaying the watershed boundaries on the maps in Figure 4.1, Figure 4.2, and Figure 4.3, respectively. In the following section, homogeneous regions are delineated based on the spatially distributed representations of

soil drainage, basin slope and basin elevation by using those percentages as similarity measures in the Ward clustering technique.

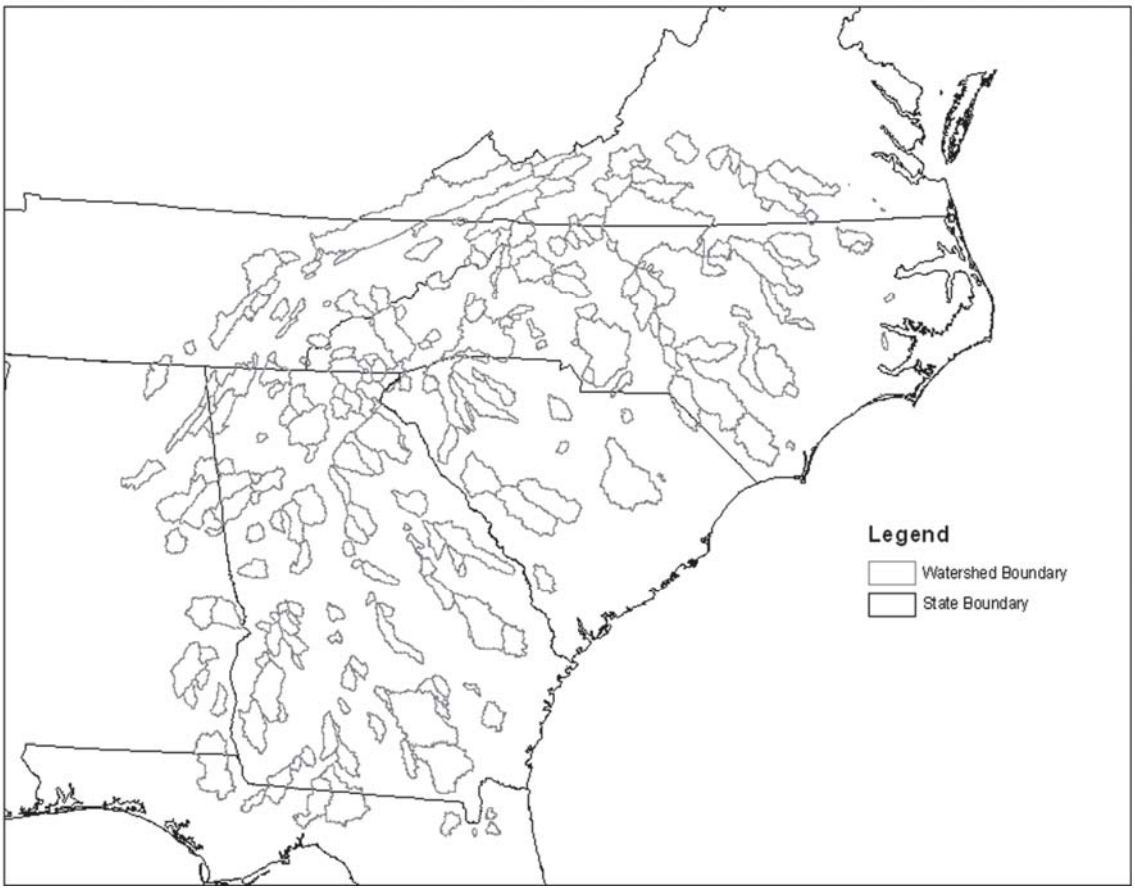


Figure 4.4: Spatial extent of watersheds corresponding to gauging stations in Figure 3.1.

4.3 Region Delineation Using Spatially Distributed Attributes

Hydrologically homogeneous regions were delineated for Sets 1 and 2 using the standardized procedure outlined in Section 3.1, wherein the attributes used in Wards clustering technique are spatially distributed values of soil drainage, basin slope, and basin elevation as defined in the preceding section. The results for each data set are discussed below.

Seven regions, as illustrated in Figure 4.5, were delineated for Set 1 using Wards clustering technique applied with the spatially distributed attributes. In general, the regions are spatially contiguous and would be adequate for visual classification of ungauged basins. Table 4.4 reports the number of sites (N) contained in each region, the corresponding value of the H-statistic computed using equation (23), and the total record length available ($\sum_{all\ i} n_i$). These metrics are presented for the original regions resulting from the Ward clustering technique, as well as for regions modified by removing sites identified as hydrologically discordant (HD) using equation (26), and for regions modified by removing only the sites which are both hydrologically and physically discordant (HPD), wherein physical discordancy is assessed using equation (54). The discordant sites are also illustrated in Figure 4.5, wherein circles represent HD sites and squares represent HPD sites; the color of the symbol corresponds to the original region within which the site in question was classified. In addition, the last row of Table 4.4 reports the average heterogeneity level (H*) computed using equation (57) as an overall measure of the ability of the delineated regions to be used for flood quantile estimation. Overall, the value of H* indicates that the delineation for Set 1 is acceptable; however, only four of the original regions are sufficiently homogeneous with H < 4. The homogeneity of the other three regions can be improved by removing HD sites; however, this is ill-advised in the interest of deriving quantile estimators for ungauged sites. Unfortunately, none of these sites are physically discordant, and thus estimators derived for Regions 2, 4, and 6 should be used with caution. It is interesting to note that when HD sites are removed from Regions 5 and 7, the total available record length within each

region drops below 500 years, and thus estimators of the 100-year event derived therein would be suspect.

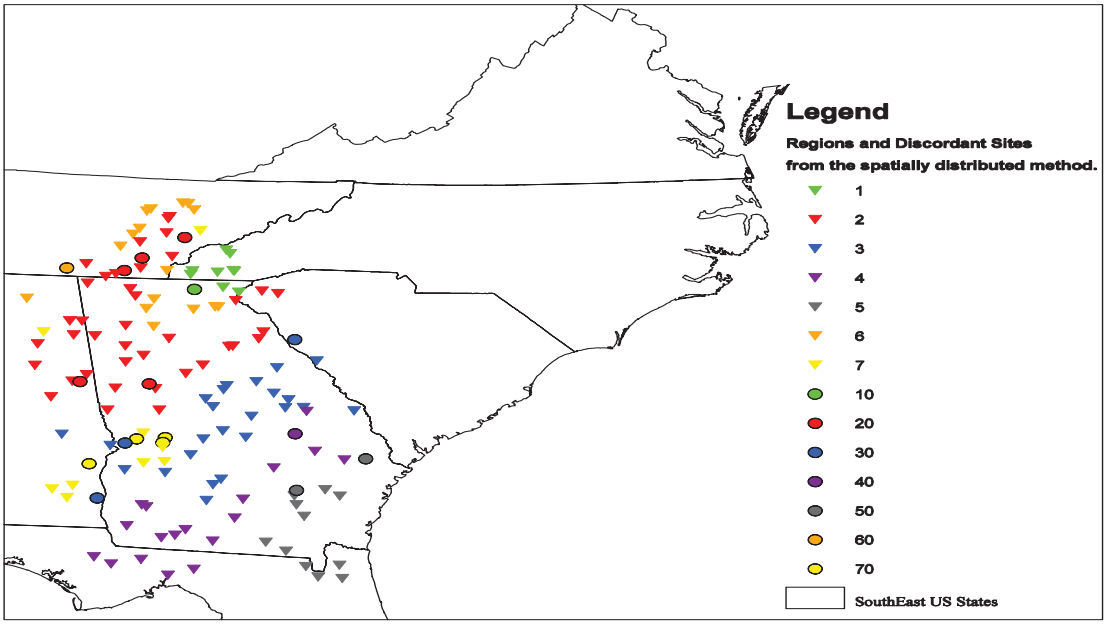


Figure 4.5: Region delineation for Set 1 obtained using Wards clustering applied to spatially distributed values of elevation, basin slope, and soil drainage. Circles represent HD sites, and squares represent HPD sites.

Table 4.4
Size and homogeneity of regions delineated using Wards clustering applied to spatially distributed values of basin elevation, basin slope, and soil drainage for sites in Set 1.

	Original Regions			**Regions without HD sites**			**Regions without HPD sites**		
	N	H	$\sum_{all\ i} n_i$	N	H	$\sum_{all\ i} n_i$	N	H	$\sum_{all\ i} n_i$
Region 1	10	2.27	727	9	2.86	667	10	2.27	727
Region 2	44	4.80	1923	39	2.83	1747	44	4.80	1923
Region 3	27	2.67	1053	24	3.39	915	27	2.67	1053
Region 4	19	4.91	829	18	4.81	798	19	4.91	829
Region 5	13	2.57	516	11	2.20	455	13	2.57	516
Region 6	17	5.55	947	16	3.57	894	17	5.55	947
Region 7	13	-0.02	540	9	0.63	402	13	-0.02	540
H*	2.9			3.1			2.9		

The same procedure was repeated for the sites contained in Set 2. Six regions were delineated as illustrated in Figure 4.6. These regions are sufficiently contiguous for visual classification of ungauged sites. Table 4.5 reports the number of sites (N) contained in each region, the corresponding values of the H-statistic and the total record length available ($\sum_{all\ i} n_i$), as well as the average heterogeneity level (H*). These metrics are presented for the original regions resulting from the Ward clustering technique, as well as for regions modified by removing all HD sites, and for regions modified by removing only HPD sites. The discordant sites are also illustrated in Figure 4.6 wherein circles represent HD sites and squares represent HPD sites. Overall, the value of H* indicates that the delineation for Set 2 is acceptable; however, two of the original regions (1 and 6) are not sufficiently homogeneous as H ≥ 4. The homogeneity of Region 1 is

reduced to 2.91 following the removal of an HPD site; however, the homogeneity of Region 6 cannot be improved as no HPD sites are identified therein.

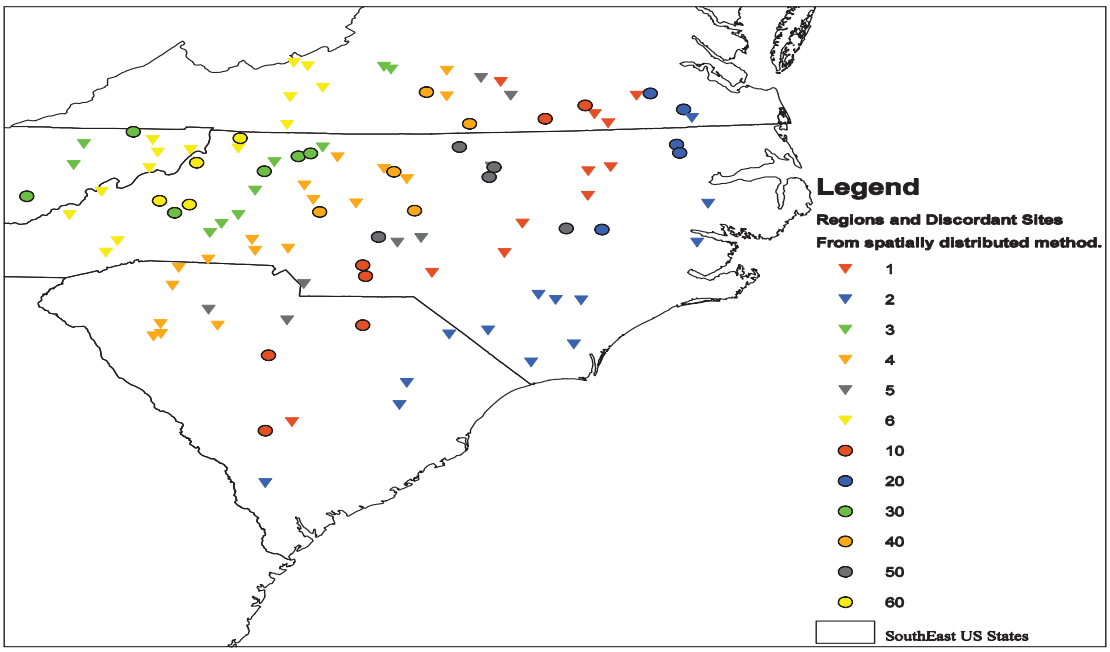


Figure 4.6: Region delineation for Set 2 obtained using Wards clustering applied to spatially distributed values of basin elevation, basin slope, and soil drainage. Circles represent HD sites, and squares represent HPD sites.

Table 4.5
Size and homogeneity of regions delineated using Wards clustering applied to spatially distributed values of basin elevation, basin slope, and soil drainage for sites in Set 2.

	Original Regions			**Regions without HD sites**			**Regions without HPD sites**		
	N	H	$\sum_{all\ i} n_i$	N	H	$\sum_{all\ i} n_i$	N	H	$\sum_{all\ i} n_i$
Region 1	18	5.00	1064	11	2.19	709	17	2.91	983
Region 2	18	1.52	880	13	3.06	660	17	1.34	854
Region 3	16	2.38	906	10	2.71	561	15	1.24	877
Region 4	23	3.75	1239	18	2.64	999	23	3.75	1239
Region 5	13	3.04	687	8	1.82	438	10	1.91	528
Region 6	18	4.14	1190	14	1.14	941	18	4.14	1190
H*		3.4			2.2			2.8	

4.4 Accuracy of Quantile Estimators for Ungauged Basins

Acceptable homogeneous regions were delineated in Chapter 3 using the standardized procedure presented therein applied with aggregated values of three key physical variables: basin slope (S_B), basin elevation (E), and soil drainage (SI). In Section 4.3, acceptable homogeneous regions were delineated in the same fashion; however, spatially distributed representations of the three key variables were employed as attributes in the Ward clustering technique. For both Sets 1 and 2, the same number of regions were delineated in each case; however, the actual sites composing each region were not necessarily the same -- overall, there was roughly 66% consistency in the region delineation.

In general, for both Sets 1 and 2, the H-statistics for regions delineated as a function of aggregated parameter values (see Table 3.10 and Table 3.15) are comparable to those of regions delineated as a function of spatially distributed variables. This suggests that regions delineated using aggregated parameter values are sufficient for the regionalization of flood data. To further justify this conclusion, the jackknife resampling procedure described in Section 3.5 is repeated herein to evaluate the accuracy of quantile estimators derived from regions delineated using spatially distributed variables; these results are subsequently compared to the results in Chapter 3 derived using aggregated parameter values. The accuracy of quantile estimators corresponding to various return periods of interest are assessed in terms of the mean square error (MSE) computed using equation (56). Figure 4.7 and Figure 4.8 illustrate the MSEs of quantile estimators derived for ungauged basins within each delineated region of Set 1 and Set 2, respectively, obtained using spatially distributed variables. Results are for either the original regions resulting directly from the Ward clustering scheme in the instance that $H < 4$, or those which have been modified by removing the HPD sites in the instance that $H \geq 4$. Additional results are tabulated in Appendix C. The highest errors obtained were approximately 17%, corresponding to estimation of the 100-year event in Region 4 of Set 1 and Region 1 of Set 2, both of which are approximately the same distance from the coastline. For Region 4 of Set 1, this large error is likely due to the high level of heterogeneity ($H = 4.91$) therein (see Table 4.4); however, larger H values do not always correlate with larger MSEs. In fact, the H-statistic computed for Region 1 of Set 2 ($H = 2.91$) is within the range of acceptable homogeneity (see Table 4.5). Further, the second

and third largest MSEs (on the order of 12%) for estimators of the 100-year event in Set 2 are observed in regions with H values less than 2.

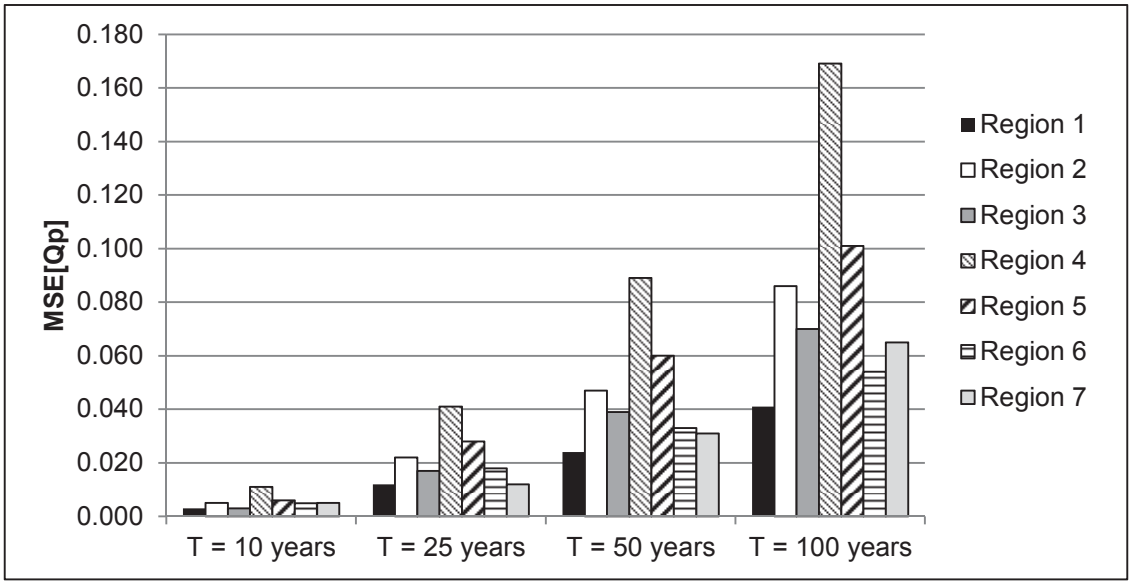


Figure 4.7: MSE of flood quantiles obtained for regions delineated in Set 1 using spatially distributed representations of basin slope, elevation, and soil drainage.

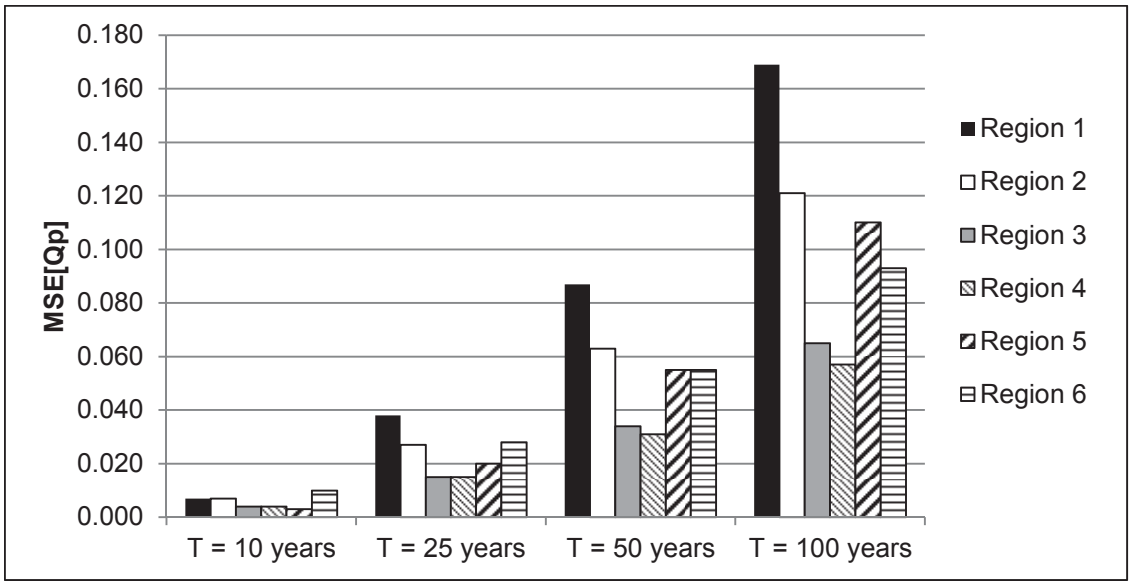


Figure 4.8: MSE of flood quantiles obtained for regions delineated in Set 2 using spatially distributed representations of basin slope, elevation, and soil drainage.

The efficiency of the quantile estimators derived using regions delineated based on spatially distributed variables relative to those derived using regions delineated using aggregated parameter values can be assessed by comparing the results in Figure 4.7 and Figure 4.8 above to those in Figure 3.13 and Figure 3.16 for Sets 1 and 2, respectively. For ease of comparison, the MSEs averaged across the regions delineated in each case are illustrated in Figure 4.9 below. For each dataset, the average MSEs obtained using spatially distributed variables are comparable to those obtained using aggregated variables. However, with respect to the accuracy reported for individual regions, the maximum error obtained when aggregated variables are employed was on the order of 13% in both datasets versus the maximum errors of 17% reported above for the spatially distributed variables. Overall, gains in the accuracy of quantile estimators for ungauged basins are not observed when regions are delineated as a function of spatially distributed variables. Thus, the physical processes that influence the flood regime appear to be adequately captured by attributes aggregated at the watershed scale.

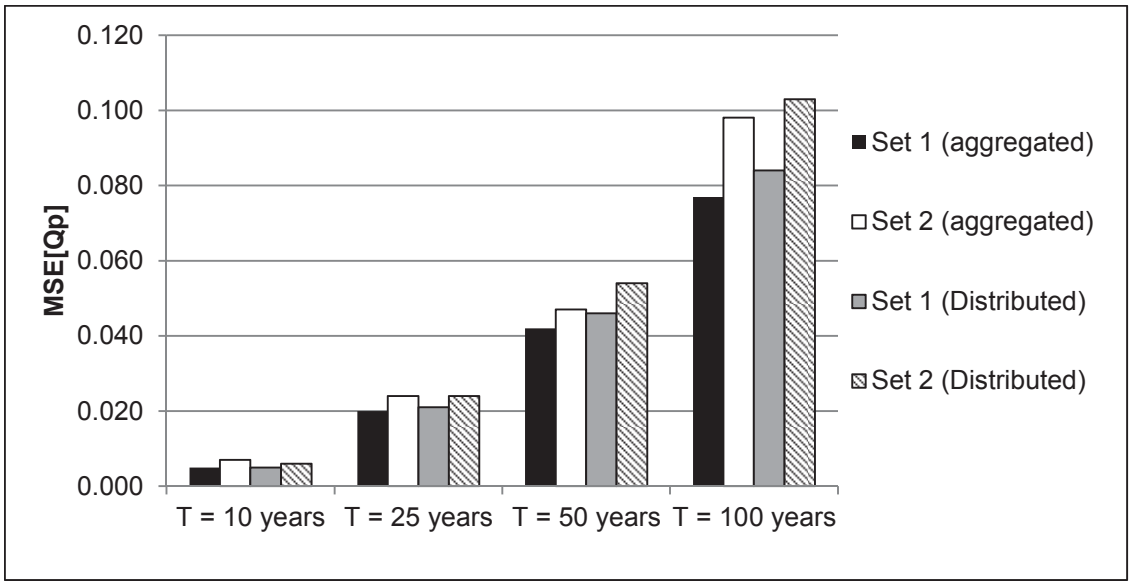


Figure 4.9: Average MSE of flood quantiles obtained across regions delineated in Sets 1 and 2 using either aggregated or spatially distributed representations of basin slope, elevation, and soil drainage.

4.5 Conclusion

This chapter presented a novel methodology to delineate regions based on spatially distributed watershed variables. Regions throughout the Southeastern U.S. were delineated as a function of spatially distributed values of soil drainage, basin slope, and basin elevation. The homogeneity of the delineated regions and the accuracy of quantile estimators derived for application at ungauged sites therein were assessed. Only marginal gains in the average heterogeneity (H*) were observed when regions were delineated using spatially distributed variables instead of aggregated values. Herein, 90m DEMs were used to obtain spatial representations of basin slope and elevation. Perhaps additional gains could be observed in areas for which complete DEMs of higher resolution are available. However, it is unclear whether these gains

would translate into increased accuracy of quantile estimators. Overall, the quantile estimators derived using regions delineated as a function of aggregated values were observed to perform slightly better than those derived as a function of spatially distributed values. Thus, although hydrologic models of individual watersheds are generally improved using spatially distributed parameters, the results presented herein do not suggest that use of physical attributes defined at scales finer than that of the watershed would be beneficial for the regionalization of flood data. This conclusion is fortunate, as simple models for the extrapolation of results to data sparse areas outside of the study region can be developed as a function of aggregated watershed variables. This type of application is investigated in Chapter 5.

Chapter 5 : Development of Quantile Estimators for Ungauged Basins in Data Sparse Areas

The Index Flood method has been praised for its effectiveness in increasing the accuracy of flood quantile estimates at gauged sites for which the available record length is short (e.g., Hosking and Wallis 1997); however, extensions to ungauged sites are rarely sought (e.g., Madsen and Rosbjerg 1997; Kjeldsen and Jones 2006). This is primarily due to the need to develop an estimate of the at-site scale parameter for ungauged sites as a function of measurable characteristics, as well as difficulty classifying ungauged sites when the delineated regions are not sufficiently contiguous. The former issue can be resolved using generalized least squares (GLS) regression models of the mean (e.g., Stedinger et al. 1993), as discussed in Section 2.2.2 herein; the latter could be resolved by using linear discriminant functions to classify ungauged sites (e.g., Chiang et al. 2002b) within regions which were appropriately delineated using physical watershed characteristics. Alternatively, quantile estimates for ungauged basins could be obtained directly as a function of watershed characteristics using GLS regression models developed for a predefined region. In the United States, flood quantile estimators for each state in the nation have been developed by the USGS using weighted least squares (WLS) or GLS regression procedures (e.g., Ries and Crouse 2002; Feaster et al. 2009; Gotvald et al. 2009; Weaver et al. 2009). However, use of regression models to obtain quantile estimates directly has a significant disadvantage in that separate models must be developed for each quantile of interest, whereas the Index Flood procedure provides the entire frequency distribution at a given site from which any quantile could be derived.

As currently practiced, both the Index Flood method and regional regression models can only be applied within the predefined area used for model development. The main goal of this chapter is to demonstrate that quantile estimators derived for data rich regions can be extrapolated to data sparse areas with similar physical characteristics which are located outside of the area used for model development. In particular, flood quantile estimates for the Gonaives watershed in Haiti are derived using relationships developed for the western portion of the Southeastern U.S. (Set 1). This is accomplished using a hybrid GLS-Index Flood approach wherein quantile estimators for ungauged sites are derived using GLS regression models of the mean to scale the regional flood distribution; ungauged watersheds external to Set 1 are classified within the regions delineated for Set 1 using linear discriminant functions. Prior to the application in Haiti, the methodology is first validated by considering the sites in the eastern portion of the Southeastern U.S. (Set 2) to be ungauged.

5.1 Estimation of the At-site Mean via GLS Regression

In Chapter 3, seven regions were delineated for Set 1 using aggregated values of the mean elevation, mean basin slope, and soil drainage index as attributes in the Ward clustering technique (see Figure 3.5); characteristics of each of these regions are provided in Table 3.10. Herein, the regions considered are either those resulting directly from the Ward clustering scheme in the instance that $H < 4$, or those which have been modified by removing the sites which are both hydrologically and physically discordant (HPD) in the

instance that H ≥ 4. For each of the seven regions, models of the following form were considered:

$$\log(\mu_i) = b_o + b_1 \log(A_i) + b_2 \log(S_{Chi}) + b_3 \log(S_{Bi}) + b_4 \log(S_{hi}) + b_5 \log(E_i) + b_6 \log(Imp_i) + b_7 \log(F_i) + b_8 \log(SI_i) + b_9 \log(Inf_i) + b_{10} \log(\mathrm{Pr}_i) + e_i \quad (58)$$

where μ_i is the mean of the flood flows at site *i* (equivalent to the first L-moment, λ_1), and e_i is the estimation error. In addition to the nine physical characteristics considered in previous chapters, the mean annual precipitation at site i (Pr_i) is investigated as a possible explanatory variable. All of the explanatory variables employed are the original untransformed values of the physiographic and meteorological characteristics.

For each region, the coefficients of the model in equation (58) were estimated using a GLS analysis (see Section 2.2.2) for which the needed estimator of the sampling covariance matrix $\mathbf{\Sigma}$ is defined in equation (51). To avoid correlation among the residuals, elements of $\mathbf{\Sigma}$ are computed using sample coefficients of variation estimated as a function of the drainage area using a separate regression analysis (see below), and lag-zero cross-correlations between sites approximated using the following smoothing function (Tasker and Stedinger 1989):

$$\rho_{ij} = \theta^{\left(\frac{d_{ij}}{\alpha d_{ij}+1}\right)} \quad (59)$$

where ρ_{ij} is the lag-zero cross-correlation between sites i and j, d_{ij} is the distance between sites i and j, and α and θ are function parameters to be estimated. Figure 5.1 illustrates the lag-zero cross-correlations plotted as a function of the distance between sites in

Region 1 of Set 1, as well as the smoothing function obtained for use therein. Smoothing functions for the remaining regions in Set 1 are presented in Appendix D.

An estimator of the sample coefficient of variation of the flood flows at site i (CV_i) appropriate for use in equation (51) is given by the model (e.g., Madsen and Rosbjerg 1997):

$$log(CV_i) = \alpha + \beta log\ (A_i) + \varepsilon_i \tag{60}$$

where A_i is the area of the watershed corresponding to site i, and ε_i is the model error. Additional explanatory variables could be used, but the area alone is sufficient for the purpose of this estimator herein. Coefficients of the model (α and β) for each region were estimated using a simple ordinary least squares (OLS) regression analysis. The models obtained for each region are presented in Table 5.1; summary statistics indicating the precision of each model are tabulated in Appendix D.

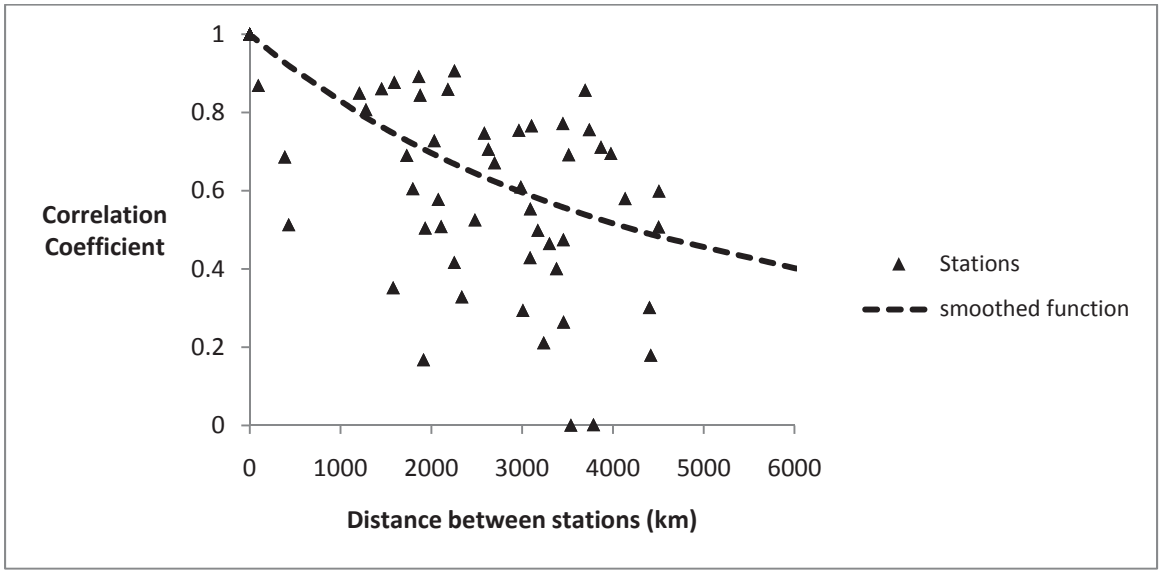


Figure 5.1: Correlation-distance smoothing function for Region 1 (α = 0.000053, θ = 0.9998).

Table 5.1
OLS regression models of the coefficient of variation derived for each region in Set 1.

Region 1	$\log(CV) = -0.9391 + 0.0547 log\ (\mathrm{A})$
Region 2	$log(CV) = -0.9128 + 0.0109 log\ (\mathrm{A})$
Region 3	$log(CV) = -0.5755 - 0.0600 log\ (\mathrm{A})$
Region 4	$log(CV) = -0.7600 - 0.0600 log\ (\mathrm{A})$
Region 5	$log(CV) = -0.8480 + 0.0027 log\ (\mathrm{A})$
Region 6	$\log(CV) = -0.6244 - 0.0093 log\ (\mathrm{A})$
Region 7	$log(CV) = -0.6430 - 0.0237 log\ (\mathrm{A})$

Using the estimators of the coefficient of variation and cross-correlation developed above to estimate the sampling covariance matrix, various forms of the model in equation (58) were fit using GLS regression. For each region, the first form of the model considered employed all 10 explanatory variables in addition to the constant term. The significance of the coefficients corresponding to each variable were assessed at the 5% level using a two-sided hypothesis test, and the GLS procedure was repeated using only the variables for which the coefficients were significant at the 5% level. This process was repeated until all coefficients remained significant. The final GLS regression models of the mean derived for each region in Set 1 are presented in Table 5.2; summary statistics indicating the precision of each model are reported in Table 5.3. Plots of the residuals and summary statistics for the coefficients for each of these models are provided in Appendix D. As is generally expected, the drainage area plays a key role in each of these models. The statistics for all of the models, except that for Region 3, indicate a reasonable level of precision, with relatively low values of model error

variance (σ_δ^2) and average variance of predication (AVP), coupled with high values of R^2_{GLS}. (See Section 2.2.2 for details regarding computation of these metrics.) In the following sections, flood quantile estimates for ungauged basins are derived via the Index Flood procedure wherein these models of the mean are used to estimate the at-site scaling factor for stations located both inside and outside of Set 1.

Table 5.2
GLS regression models of the at-site mean derived for each region in Set 1.

Region 1	$\log(\mu) = -3.610 + 0.822\log(\text{A}) + 2.644\log(\text{F}) + 1.336\log(\text{Inf})$
Region 2	$log(\mu) = 6.150 + 0.744\log(\text{A}) - 0.466\log(\text{Sh}) - 7.260\log(\text{SI})$
Region 3	$\log(\mu) = 0.899 + 0.954\log(\text{A}) + 0.710\log(\text{S}_{Ch}) + 0.124\log(\text{Imp}) - 0.073\log(\text{F})$
Region 4	$\log(\mu) = 2.061 + 0.691\log(\text{A}) - 0.338\log(\text{Sh}) + 1.035\log(\text{Inf})$
Region 5	$\log(\mu) = -4.386 + 0.568\log(\text{A}) + 2.289\log(\text{Pr}) + 1.6891\log(\text{F})$
Region 6	$\log(\mu) = -8.734 + 0.5861\log(\text{A}) + 0.592\log(\text{S}_\text{B}) + 5.772\log(\text{Pr}) + 2.061\log(\text{Inf})$
Region 7	$\log(\mu) = 5.840 + 0.401\log(\text{A}) - 0.873\log(\text{Sh}) - 10.522\log(\text{SI}) + 9.212\log(\text{Inf})$

Table 5.3
Summary statistics for GLS regression models of the at-site mean derived for each region in Set 1.

	AVP	σ_{δ}^{2}	R^2_{GLS}
Region 1	0.049	0.0018	0.986
Region 2	0.1629	0.0220	0.823
Region 3	0.2398	0.0467	0.640
Region 4	0.1063	0.0097	0.908
Region 5	0.0978	0.0079	0.867
Region 6	0.1280	0.0138	0.901
Region 7	0.1507	0.0160	0.826

5.2 Flood Quantile Estimation at Ungauged Sites within the Predefined Study Area

In order to assess the accuracy of the GLS models of the mean in the context of flood quantile estimation via the Index Flood method, the jackknife resampling procedure presented in Section 3.5 is applied herein using data for sites within the seven regions delineated for Set 1. For a given region, the accuracy of quantile estimators corresponding to various return periods of interest are assessed in terms of the Mean Square Error (MSE) computed using equation (56) wherein the at-site estimator is derived from a GEV distribution fit to the available flood record using L-moments, and the regional estimator is derived using a hybrid GLS-Index Flood (GLS-IF) approach assuming that the site in question is ungauged. The GLS-IF approach utilizes the models in Table 5.2 to estimate the at-site mean for an ungauged site; this estimate is then used to

scale the regional frequency distribution obtained via the Index Flood method in order to derive flood quantile estimates for the site in question.

Table 5.4 reports the MSEs of quantile estimators derived for ungauged sites within each region of Set 1 using the GLS-IF approach. The largest MSE obtained is nearly 50%, corresponding to estimation of the 100-year event in Region 3. This large error is likely due to the poor precision of the GLS model for the mean in that region; in general, the MSEs for a given return period tend to increase with the AVP of the GLS regression models. Relatively large MSEs are also obtained in Regions 2, 6, and 7, wherein the values of the H-statistics were also reasonably large (5.06, 3.70, and 4.95, respectively, as reported in Table 3.10). Overall, due to additional errors resulting from estimation of the mean, the GLS-IF model is less precise than use of the Index Flood procedure alone (see Figure 3.13 for comparison). However, the MSEs reported in Table 5.4 are generally acceptable when compared to the results of other regionalization studies. For instance, Malekinezhad et al. (2011) report average root mean square errors for various return periods in the range of 0.13 to 0.53, corresponding to MSEs of 0.02 to 0.28. Similarly, Shu and Ouarda (2009) report average root mean square errors ranging from 0.53 to 0.57 (corresponding to MSEs of 0.28-0.32) for estimation of the 10-year event, 0.58 to 0.76 (0.34-0.58) for the 50-year event, and 0.64 to 0.75 (0.41-0.56) for the 100-year event. Therefore, the GLS-IF approach is suitable for estimation of flood quantiles at ungauged sites within the study area used for model development when the GLS regression model for the mean can be derived with adequate precision.

Table 5.4
MSE of flood quantiles obtained using the GLS-IF approach for sites within Set 1.

	T = 10 years (p = 0.90)	T = 25 years (p = 0.96)	T = 50 years (p = 0.98)	T = 100 years (p = 0.99)
Region 1	0.0071	0.0067	0.0089	0.0137
Region 2	0.1560	0.1941	0.2331	0.2837
Region 3	0.7932	0.6995	0.5886	0.4988
Region 4	0.0738	0.0822	0.0945	0.1167
Region 5	0.0902	0.1088	0.1379	0.1877
Region 6	0.1302	0.2027	0.2783	0.3819
Region 7	0.1111	0.1668	0.2217	0.2941
Average	**0.1945**	**0.2087**	**0.2233**	**0.2538**

5.3 Extension of GLS-IF Model to Alternate Areas

Results presented above verify that application of the Index Flood procedure using GLS regression models of the mean to estimate the scale factor for ungauged sites is appropriate within the study area used for model development. However, a primary goal of this dissertation is to demonstrate that the relationships derived in a data rich study area can successfully be extrapolated to sites in areas external to that used for model development. In the sections below, the extrapolation to sites of similar physical composition is first validated using gauged sites in a data rich region (i.e., Set 2), and then an example of the extrapolation to a data sparse area is provided using data for the Gonaives watershed in Haiti.

5.3.1 Validation via Extension to Data Rich Areas

Results presented in Chapter 3 indicate that hydrologically homogeneous regions can be successfully delineated throughout the Southeastern U.S. using three key physical characteristics: mean elevation, mean basin slope, and soil drainage. Therefore, it should be possible to successfully classify sites in Set 2 within one of the seven regions delineated for Set 1, and to subsequently derive flood quantiles for the sites in Set 2 using the GLS-IF approach.

In the following analyses, the watersheds in Set 2 are considered ungauged and are classified within regions in Set 1 using linear discriminant functions that differentiate between those regions as a function of watershed characteristics. The values of the nine physical variables (normalized and standardized) for the sites contained in each of the regions in Set 1 define seven sets of variables. As there are seven sets of variables, six discriminant functions are sufficient to differentiate between the regions. Table 5.5 reports the coefficients of the linear discriminant (LD) functions corresponding to each of the physical variables; the correlation coefficients indicating the relative importance of each physical variable are provided in Table 5.6. The first discriminant function (LD_1) is the most powerful, explaining 72% of the differences among the regions; as expected, it is most influenced by the basin slope, mean elevation, and soil drainage.

Table 5.5
Coefficients of the normalized and standardized physical variables in the discriminant functions differentiating between regions in Set 1.

Physical Characteristic	**LD_1**	**LD_2**	**LD_3**	**LD_4**	**LD_5**	**LD_6**
Drainage Area	0.310	-0.246	4.380	-0.213	2.176	1.369
Main Channel Slope	0.158	0.437	5.584	0.334	4.842	3.591
Basin Slope	3.796	-12.53	4.561	9.191	-0.385	6.796
Basin Shape	1.072	-0.204	1.156	1.576	0.523	-3.705
Basin Elevation	17.93	11.95	-0.784	-3.103	-1.953	-3.035
% Impervious Cover	1.063	1.011	-1.406	3.827	2.494	1.234
% Forest Cover	-0.972	-1.679	1.105	-4.726	-4.017	-3.039
Soil Drainage Index	-1.356	4.933	6.529	3.144	-8.066	5.854
Infiltration Index	0.723	2.084	2.592	1.872	2.139	-2.695

Table 5.6
Correlation coefficients resulting from LDA applied to the normalized and standardized physical variables for regions in Set 1.

Physical Characteristic	**LD_1**	**LD_2**	**LD_3**	**LD_4**	**LD_5**	**LD_6**
Drainage Area	-0.200	-0.107	0.260	0.140	0.034	-0.252
Main Channel Slope	0.306	0.024	**0.466**	**-0.491**	**0.386**	0.303
Basin Slope	**0.784**	**-0.592**	0.100	0.131	-0.068	0.018
Basin Shape	-0.035	-0.215	0.107	0.396	-0.012	**-0.557**
Basin Elevation	**0.994**	0.080	0.001	-0.050	-0.022	-0.009
% Impervious Cover	0.007	0.260	-0.167	**0.421**	-0.071	-0.109
% Forest Cover	0.311	**-0.465**	0.373	-0.284	-0.254	-0.199
Soil Drainage Index	**-0.603**	**0.630**	0.225	0.209	**-0.344**	0.161
Infiltration Index	-0.326	0.374	**0.527**	0.337	0.070	-0.340
% Variance Explained:	0.7225	0.2093	0.0401	0.0229	0.0050	0.0002

As LD_1 explains the majority of the differences between the regions in Set 1, this function will be used to classify the watersheds in Set 2 within one of the seven regions in Set 1. This is achieved by calculating values of the LD_1 function corresponding to each site in Set 2, denoted $LD_1(i)$. In addition, values of the LD_1 function are computed using the mean values of the physical characteristics for each region in Set 1, denoted $\overline{LD_1}(k)$ for k = 1, …, 7. Each site from Set 2 is then allocated to the region in Set 1 for which $LD_1(i)$ and $\overline{LD_1}(k)$ are closest. Table 5.7 reports the number of sites from Set 2 allocated to each of the regions in Set 1. This type of allocation rule using discriminant analysis was previously employed by Chiang et al. (2002[b]). A confusion matrix is commonly used to assess the error rate of a proposed classification scheme (e.g., Johnson and Wichern 2007). Table 5.8 contains the confusion matrix for the classification scheme employed herein. Elements of the matrix reflect the fraction of sites in Set 1 that are misclassified using the proposed allocation metric based on LD_1. For example, Region 5 should contain 20 sites, however, the metric based on LD_1 misclassifies 2 sites as belonging in Region 6 and 4 sites as belonging in Region 7. Overall, the classification scheme works well, except in Regions 5 and 7. Thus, results derived herein for the latter regions should be viewed with caution. The success of the classification of ungauged sites could be improved by retaining LD_2 for use in the classification scheme (see Appendix E); however, this alternative classification scheme is not considered in the subsequent analyses presented herein.

Table 5.7
Number of sites (N) in Set 2 allocated to each region in Set 1 using LD_1.

	N
Region 1	19
Region 2	16
Region 3	14
Region 4	17
Region 5	17
Region 6	19
Region 7	4

Table 5.8
Confusion matrix for the classification scheme based on LD_1.

	Region						
Actual\Prediction	1	2	3	4	5	6	7
Region 1	12	0	0	0	0	0	0
Region 2	0	22	0	0	0	0	0
Region 3	0	0	22	0	0	0	0
Region 4	0	0	0	26	0	0	0
Region 5	0	0	0	0	14	2	4
Region 6	0	0	3	0	0	24	0
Region 7	0	0	0	0	10	1	3
Prediction Accuracy (%):	100	100	88	100	58	89	43

The efficiency of regional quantile models extrapolated from Set 1 for application within Set 2 was assessed in two phases. In Phase I, the Index Flood relationships derived within Set 1 were scaled for application at sites in Set 2 using the actual at-site

mean as a scale parameter in order to further validate the use of mean elevation, basin slope and soil drainage to characterize the flood regime throughout the Southeastern U.S. In this way, the sites in Set 2 are considered gauged, and thus for the extrapolation to be successful, the quantile estimators derived in this way should be as efficient as those derived for sites contained in Chapter 3 (see Figure 3.16). In Phase II, sites in Set 2 were considered ungauged and the GLS-IF approach was used to derive flood quantile estimates wherein the Index Flood models derived for Set 1 were scaled to the site in question in Set 2 using the appropriate GLS model of the mean based on physical characteristics. In this case, if the extrapolation is successful, then the results should be comparable to those reported in Table 5.4, corresponding to application of the GLS-IF approach within the area used for model development (i.e., Set 1). This is the first time regional models have been extrapolated to sites located outside of the study area used for model development. Results of jackknife resampling procedures used to assess the efficiency of quantile estimators derived in each phase are presented below.

In Phase I, wherein sites in Set 2 are considered gauged, the accuracy of quantile estimators corresponding to various return periods of interest is assessed in terms of the MSE computed using equation (56) where the at-site estimator is derived from a GEV distribution fit to the available flood record using L-moments, and the regional estimator is derived using the Index Flood model developed for the region in Set 1 into which the site from Set 2 was classified; the regional frequency distribution is scaled using the actual at-site mean for the site in Set 2. Table 5.9 reports the MSEs of quantile estimators derived for gauged sites in Set 2 allocated to each region of Set 1. The largest MSE

obtained is nearly 30% corresponding to estimation of the 100-year event for sites allocated to Region 7; however, only 4 sites from Set 2 were allocated therein. Omitting the results for Region 7, the average error is on the order of 7%, which is consistent with the performance of the Index Flood method applied directly to Set 2 (see Figure 3.16). Therefore, these results indicate that the Index Flood models developed for Set 1 can successfully be extrapolated to sites in Set 2. Thus, when the watersheds in Set 2 are considered ungauged, only an estimate of the at-site mean should be necessary to scale the regional quantile model. This is considered in Phase II below.

In Phase II, wherein sites in Set 2 are considered ungauged, the accuracy of quantile estimators derived using the GLS-IF approach to extrapolate results from Set 1 is assessed in terms of the MSE. Herein, the at-site estimator is derived from a GEV distribution fit to the available flood record using L-moments, and the regional estimator is derived using the Index Flood model developed for the region in Set 1 into which the site from Set 2 was classified; the regional frequency distribution is scaled using an estimator of the mean derived for the site of interest in Set 2 using the appropriate GLS regression model from Table 5.2. Table 5.10 reports the MSEs of quantile estimators derived in this way for sites in Set 2 allocated to each region of Set 1. Again, the largest MSE obtained corresponds to estimation of the 100-year event for sites allocated to Region 7, but with only 4 sites from Set 2 allocated therein, these results are suspect. The average error is cut in half when the results for Region 7 are omitted; however, the MSEs in Regions 2 and 5 are also unacceptable. For Region 5, this may be partially due to the high classification error reported in Table 5.8.

Table 5.9

MSE of quantiles estimated for sites in Set 2 using the Index Flood model for the corresponding region in Set 1 scaled by the at-site mean computed using the available record at the site of interest (Phase I).

	T = 10 years (p = 0.90)	**T = 25 years (p = 0.96)**	**T = 50 years (p = 0.98)**	**T = 100 years (p = 0.99)**
Region 1	0.0031	0.0016	0.0113	0.0310
Region 2	0.0041	0.0065	0.0134	0.0265
Region 3	0.0277	0.0539	0.0933	0.1615
Region 4	0.0114	0.0137	0.0206	0.0355
Region 5	0.0157	0.0079	0.0113	0.0285
Region 6	0.0341	0.0619	0.0962	0.1523
Region 7	0.0293	0.0852	0.1661	0.2999
Average	**0.0179**	**0.0329**	**0.0589**	**0.1050**

Table 5.10

MSE of quantiles estimated for sites in Set 2 using the GLS-IF approach wherein the GLS estimator of the mean derived for the site of interest is used to scale the Index Flood model for the corresponding region in Set 1 (Phase II).

	T = 10 years (p = 0.90)	**T = 25 years (p = 0.96)**	**T = 50 years (p = 0.98)**	**T = 100 years (p = 0.99)**
Region 1	0.1927	0.2141	0.2407	0.2763
Region 2	1.0213	1.0117	1.0088	1.0284
Region 3	0.2914	0.2681	0.2516	0.2445
Region 4	0.0908	0.0758	0.0666	0.0649
Region 5	3.7972	4.5493	5.2547	6.1203
Region 6	0.1394	0.1502	0.1678	0.2024
Region 7	5.4195	6.8575	8.2561	10.126
Average	**1.5646**	**1.8752**	**2.1780**	**2.5804**

Kjeldsen and Jones (2006) demonstrated that the uncertainty in quantile estimates derived using the Index Flood model increased by eight times when applied to ungauged sites within the area used for model development. Although the errors reported in Table 5.10 are within the limits of those observed by Kjeldsen and Jones (2006), it is likely that the estimators for sites external to the original study area could be improved by accounting for differences in precipitation. Table 5.11 reports summary statistics for the mean annual precipitation averaged across sites within each region delineated for Set 1, as well as for the sites from Set 2 allocated to each of those regions. Wilcoxon-Mann-Whitney tests were performed to assess the difference in the medians of the sites from Set 1 versus Set 2 contained in each region. These tests indicate that the precipitation values of the two groups are significantly different at the 5% level. (Details of these tests are provided in Appendix E, as well as results of tests and summary statistics compiled for each of the nine physical variables.)

The impact of differences in precipitation on the GLS-IF approach for extrapolation to alternate areas was assessed by repeating the jackknife resampling analysis employed in Phase II above. Herein, the regional quantile estimates are again derived using the GLS-IF approach, but prior to utilizing the Index Flood regional distribution, the estimator of the mean derived using the GLS regression model for the site of interest in Set 2 is scaled by the ratio of the at-site precipitation to the regional average precipitation of the corresponding region of Set 1; this approach is denoted GLS-IF(P). Table 5.12 reports the MSEs of quantile estimators derived in this way for sites in Set 2 allocated to each region of Set 1. Overall, accounting for differences in

precipitation improves the efficiency of quantile estimators obtained by extrapolating regional flood distributions from Set 1 for application in Set 2. Again, the error observed in Region 7 is quite high; however, four sites do not provide a strong basis for evaluation. The large error in Region 5 can be explained by the classification errors reported in Table 5.8. Omitting these regions, the average error is reduced to 22.5%, which is in line with the range of values reported in Table 5.4 for application of the GLS-IF approach within the area used for model development. These results indicate that regional flood distributions derived within data rich areas can be successfully extrapolated to alternate areas using a GLS-IF approach properly scaled to account for differences in precipitation, provided the sites in question can be successfully classified within a delineated region.

Table 5.11
Summary statistics for the mean annual precipitation (inches) for the seven regions delineated for Set 1 versus that of sites from Set 2 allocated to those regions.

	Region						
	1	**2**	**3**	**4**	**5**	**6**	**7**
				Set 1			
Minimum	59.3	50.2	46.8	50.0	47.1	46.2	47.3
Maximum	81.8	70.3	61.2	57.8	60.0	54.5	57.0
Average	70.0	61.1	49.5	53.5	54.5	49.4	53.3
Std. Dev.	6.17	4.5	3.2	2.1	2.6	2.4	3.0
				Set 2			
Minimum	40.1	41.7	45.1	44.9	45.6	47.6	48.4
Maximum	72.7	56.4	52.4	49.0	56.3	49.5	51.7
Average	53.8	48.3	47.1	46.4	48.5	48.4	50.1
Std. Dev.	8.0	4.1	1.6	1.4	2.5	0.6	2.4

Table 5.12
MSE of quantiles estimated for sites in Set 2 using GLS-IF(P) approach wherein the GLS estimator of the mean derived for the site of interest is scaled by the ratio of the at-site precipitation to the mean precipitation of the region in Set 1 within which it is allocated.

	T = 10 years (p = 0.90)	**T = 25 years (p = 0.96)**	**T = 50 years (p = 0.98)**	**T = 100 years (p = 0.99)**
Region 1	0.2267	0.2467	0.2670	0.2906
Region 2	0.4051	0.3919	0.3865	0.3943
Region 3	0.2819	0.2554	0.2368	0.2262
Region 4	0.0736	0.0601	0.0536	0.0536
Region 5	2.4586	2.9798	3.4727	4.0788
Region 6	0.1174	0.1210	0.1340	0.1624
Region 7	4.1176	5.2525	6.3718	7.8813
Average	1.0972	1.3290	1.5600	1.8696

5.3.2 Estimation at Ungauged Sites Located in Haiti

The Gonaives basin in Haiti is used herein to demonstrate application of the GLS-IF(P) approach to extrapolate regional flood distributions from data rich areas to data scarce locations. This application is an example of a more extreme extrapolation than that considered above for sites in Set 2, as the Gonaives watershed is located in an area for which the climate is vastly different from that of the area used for model development (i.e., Set 1).

The Gonaives watershed is located in the Northeastern region of Haiti. It is frequently in the path of tropical storms and hurricanes, and has an average annual precipitation of 46.1 inches. Estimating flood quantiles in Haiti is difficult due to lack of

streamflow data; therefore, previous studies of the watershed used rainfall-runoff models to estimate flood quantiles (e.g., Brandimarte et al. 2010; Ilorme et al. 2011). Available physical characteristics for the watershed obtained from Ilorme et al. (2011) are reported in Table 5.13. The table reports the original values of the physical characteristics, as well as values which have been normalized and standardized (or transformed) for use in the linear discriminant functions defined for Set 1 in Table 5.5. Evaluation of LD_1 using the transformed values indicates that the Gonaives watershed should be allocated to Region 4 of Set 1.

Table 5.13
Watershed Characteristics of Gonaives, Haiti.

Physical Characteristic	**Original Value**	**Transformed Value**
Drainage Area	212	2.33
Main Channel Slope	198	3.27
Basin Slope	21.0	5.07
Basin Shape	4.82	1.25
Basin Elevation	1406	10.6
% Impervious Cover	2.0	0.745
% Forest Cover	3.0	1.75
Soil Drainage Index	4.0	0.328
Infiltration Index	3.0	0.295

The at-site mean for Gonaives was estimated to be 120.8 m^3/s (4266 ft^3/s) using the GLS regression model developed for Region 4 (Table 5.2). This value was subsequently scaled by precipitation and used to scale the Index Flood model developed for Region 4 in Set 1. Table 5.14 reports values of the 10-, 25-, 50-, and 100-year events

derived for Gonaives. The estimate of the 100-year flood is about four times lower than that estimated by Ilorme et al. (2011) using rainfall-runoff modeling (1191m^3/s). As the true quantile is unknown, it is difficult to know which estimate is more accurate. Previous studies have demonstrated that, when employed within the area used for model development, regional flood frequency analyses often provide more accurate flood quantiles than hydrologic modeling (e.g., USACE, 1994). However, additional analyses are needed to confirm that the same holds true when regional flood distributions are extrapolated to areas outside of that used for model development, as is the case considered here.

Table 5.14
Flood quantiles derived for Gonaives watershed via the GLS-IF(P) approach.

	T = 10 years (p = 0.90)	T = 25 years (p = 0.96)	T = 50 years (p = 0.98)	T = 100 years (p = 0.99)
Flow (m^3/s)	187.1	215.8	307.4	369.9

5.4 Conclusion

This chapter presented a methodology by which regional flood distributions derived for data rich areas can be successfully extrapolated to basins of similar physical composition located beyond the extent of the study area used for model development. Results presented herein demonstrate that GLS regression models can provide sufficient estimators of the at-site scale (or index-flood) parameters for ungauged basins within the predefined study area; however, additional measures must be taken when the regional flood distribution is extrapolated to ungauged basins outside of the study area. In

particular, a transfer function based on precipitation must be applied to appropriately scale flood magnitudes for areas where the mean annual precipitation is significantly different from that in the study area. Overall, the success of the extrapolation is dependent upon the accuracy of the GLS model of the mean, and the ability to appropriately classify the site in question within a delineated region.

In addition, an example of the extrapolation of results from the data rich Southeastern portion of the U.S. to the data scarce Gonaives watershed of Haiti was presented. While such an extrapolation is certainly possible using the methodology presented herein, additional analyses are needed to confirm that the results obtained are in fact reasonable when regional flood distributions are extrapolated to areas in a significantly different climate. It is highly likely that additional scale parameters are needed to appropriately scale flood magnitudes when extrapolating the regional flood distribution to basins within alternative climate zones.

Chapter 6 : Conclusion

Regional flood frequency techniques, such as the Index Flood method and regional regression procedures, are commonly used to estimate flood quantiles when either flood data for the basin under study is unavailable or the record length at an individual gauging station is insufficient for reliable analyses. Successful derivation of site specific flood quantile estimates using regional models is predicated on the assumption that similarity in the flood regime (hydrological homogeneity) is indicated by similarity in physical characteristics aggregated at the watershed level. Unfortunately, the delineation of hydrologically homogeneous regions is highly subjective and is dependent on the physical similarity measures and classification techniques employed. In addition, as currently practiced, regional flood frequency models are only applied within the area used for model development; no procedures currently exist to extrapolate flood statistic-basin characteristic relationships from regions which are relatively data rich to data scarce regions where flood quantile estimates are sorely needed.

In order to improve quantile estimates for ungauged sites located in both data rich and data scarce areas, this dissertation presented simple rules to delineate hydrologically homogeneous regions based on the physical characteristics which are the most relevant indicators of extreme hydrologic response, and to extrapolate flood statistic-basin characteristic relationships to data sparse areas located beyond the extent of the area used for model development. Herein, use of the Index Flood method for quantile estimation at ungauged sites was investigated because once the regional flood distribution is properly scaled to the site in question, the Index Flood procedure provides the entire frequency

distribution at a given site from which any quantile could be derived, whereas separate models must be developed for each quantile of interest when regression procedures are employed.

Chapter 3 presented a standardized procedure for the delineation of hydrologically homogeneous regions. Key aspects of this procedure are a new statistical metric to identify physically discordant sites, and the identification of an appropriate set of physically based measures of extreme hydrologic response for use in region delineation. Results presented herein demonstrate that mean basin elevation, mean basin slope, and soil drainage are appropriate for regionalization of flood data throughout the Southeastern U.S. Use of these characteristics as similarity measures in the proposed approach for region delineation yields regions which are more homogeneous and more efficient for quantile estimation at ungauged sites via the Index Flood method than regions delineated using alternative physically-based procedures typically employed in practice. These key physical characteristics are also shown to be efficient for region delineation and quantile development in alternative study areas composed of watersheds with statistically different physical composition.

Previous studies have demonstrated that hydrologic models of individual watersheds are generally improved by using spatially distributed parameters, and thus it was of interest to see if additional gains in the accuracy of quantile estimators derived via the Index Flood method could be achieved by using more precise descriptions of the key physical variables. In Chapter 4, regions throughout the Southeastern U.S. were delineated as a function of spatially distributed values of soil drainage, basin slope, and

basin elevation. Overall, marginal gains in homogeneity and a slight decrease in the performance of quantile estimators for ungauged sites were observed when regions were delineated using spatially distributed variables instead of aggregated values. Thus, the use of aggregated values of key watershed characteristics is sufficient for the regionalization of flood data. This finding is fortunate, as simple models for the extrapolation of results to data sparse areas outside of the study region can then be developed.

Chapter 5 presented a methodology by which regional flood distributions derived for data rich areas can be successfully extrapolated to basins of similar physical composition located beyond the extent of the study area used for model development. Results presented herein demonstrate that GLS regression models can provide sufficient estimators of the at-site scale (or index-flood) parameters for ungauged basins within the predefined study area; however, additional measures must be taken when the regional flood distribution is extrapolated to ungauged basins outside of the study area. In particular, a transfer function based on precipitation must be applied to appropriately scale flood magnitudes for areas where the mean annual precipitation is significantly different from that in the study area. Overall, the success of the extrapolation is dependent upon the accuracy of the GLS model of the mean, and the ability to appropriately classify the site in question within a delineated region.

Overall, the research presented herein provides a critical contribution to the profession as appropriate methods for flood quantile estimation in ungauged basins are needed throughout the world, particularly in data sparse areas. An example of the

extrapolation of regional flood distributions from the data rich Southeastern U.S. to the data scarce Gonaives watershed of Haiti was presented in Chapter 5. While such an extrapolation is certainly possible using the physically-based methodology presented herein, coupled with the increasing availability of remotely sensed data, additional analyses are needed to confirm that the quantile estimates obtained are in fact reasonable when regional flood distributions are extrapolated to areas in a significantly different climate. It is highly likely that additional scale parameters are needed to appropriately scale flood magnitudes when extrapolating the regional flood distribution to basins within alternative climate zones.

In general, the ability to successfully estimate flood quantiles at ungauged sites as a function of physical attributes which are easily computed at the watershed scale using GIS and remotely sensed data makes regional flood frequency techniques quite attractive, because the development of these models as proposed herein is less time consuming and data intensive than alternatives such as watershed specific rainfall-runoff models with spatially distributed parameters. Further, previous studies have demonstrated that, when employed within the area used for model development, regional flood frequency analyses are more efficient for flood quantile estimation than hydrologic models. However, additional analyses are needed to confirm that the same holds true when regional flood distributions are extrapolated to areas outside of that used for model development.

Future work will include the development of rainfall-runoff models for select sites in Set 2 in order to assess the relative accuracy of hydrologic models versus the GLS-IF models extrapolated beyond the extent of the area used for model development. In

addition, the possible impacts of variability in precipitation across individual watersheds on quantile estimators will be evaluated. Alternative transfer functions based on factors such as the average annual 24-hour maximum precipitation event and the 24-hour, N-year storm event where N is a function of the drainage area will also be considered. Impacts of the underlying geology on the scaling of regional quantiles to individual watersheds will also be investigated. And, to ensure that the methodology and key watershed characteristics are truly applicable in range of geographic locations, the analyses presented herein will be repeated for alternate study areas such as the Midwestern and Pacific Northwestern regions of the United States.

References

Abbott MB, Bathurst JC, Cunge JA, Oconnell PE, Rasmussen J. 1986. An Introduction to the European Hydrological System - Systeme Hydrologique Europeen, She .1. History and Philosophy of a Physically-Based, Distributed Modeling System. *Journal of Hydrology*. 87(1-2):45-59.

Ajami NK, Gupta H, Wagener T, Sorooshian S. 2004. Calibration of a semi-distributed hydrologic model for streamflow estimation along a river system. *Journal of Hydrology*. 298(1-4):112-135.

Alcazar J, Palau A. 2010. Establishing environmental flow regimes in a Mediterranean watershed based on a regional classification. *Journal of Hydrology*. 388(1-2):41-51.

Baeriswyl PA, Rebetez M. 1997. Regionalization of precipitation in Switzerland by means of principal component analysis. *Theoretical and Applied Climatology*. 58(1-2):31-41.

Bates BC, Rahman A, Mein RG, Weinmann PW. 1998. Climatic and physical factors that influence the homogeneity of regional floods in southeastern Australia. *Water Resources Research*. 34(12):3369-3381.

Bertoldi L, Massironi M, Visona D, Carosi R, Montomoli C, Gubert F, Naletto G, Pelizzo MG. 2011. Mapping the Buraburi granite in the Himalaya of Western Nepal: Remote sensing analysis in a collisional belt with vegetation cover and extreme variation of topography. *Remote Sensing of Environment*. 115(5):1129-1144.

Beven KJ, Kirkby MJ, Schofield N, Tagg AF. 1984. Testing a Physically-Based Flood Forecasting-Model (Topmodel) for 3 UK Catchments. *Journal of Hydrology*. 69(1-4):119-143.

Bhaskar NR, Oconnor CA. 1989. Comparison of Method of Residuals and Cluster-Analysis for Flood Regionalization. *Journal of Water Resources Planning and Management*. 115(6):793-808.

Box GEP, Cox DR. 1982. An Analysis of Transformations Revisited, Rebutted. *Journal of the American Statistical Association*. 77(377):209-210.

Boyle DP, Gupta HV, Sorooshian S, Koren V, Zhang ZY, Smith M. 2001. Toward improved streamflow forecasts: Value of semidistributed modeling. *Water Resources Research*. 37(11):2749-2759.

Di Baldassarre G, Brandimarte L, Brath A, Castellarin A. 2009. Isla Hispaniola: A trans-boundary flood risk mitigation plan. *Physics and Chemistry of the Earth*. 34(4-5):209-218.

Brink AB, Eva HD. 2009. Monitoring 25 years of land cover change dynamics in Africa: A sample based remote sensing approach. *Applied Geography*. 29(4):501-512.

Burn DH. 1990. Evaluation of Regional Flood Frequency-Analysis with a Region of Influence Approach. *Water Resources Research*. 26(10):2257-2265.

Burn DH. 1997. Catchment similarity for regional flood frequency analysis using seasonality measures. *Journal of Hydrology*. 202(1-4):212-230.

Burn DH, Zrinji Z, Kowalchuk M. 1997. Regionalization of Catchments for Regional Flood Frequency Analysis. *Journal of Hydrologic Engineering*. 2(2):76-82.

Buttle JM, Eimers MC. 2009. Scaling and physiographic controls on streamflow behaviour on the Precambrian Shield, south-central Ontario. *Journal of Hydrology*. 374(3-4):360-372.

Butts C. T. 2009. yacca: Yet Another Canonical Correlation Analysis Package. R package version 1.1.

Castellarin A, Burn DH, Brath A. 2001. Assessing the effectiveness of hydrological similarity measures for flood frequency analysis. *Journal of Hydrology*. 241(3-4):270-285.

Chebana F, Ouarda TBMJ. 2008. Depth and homogeneity in regional flood frequency analysis. *Water Resources Research*. 44(11).

Chiang SM, Tsay TK, Nix SJ. 2002. Hydrologic regionalization of watersheds. I: Methodology development. *Journal of Water Resources Planning and Management*. 128(1):3-11.

Chiang SM, Tsay TK, Nix SJ. 2002b. Hydrologic regionalization of watersheds. II: Applications. *Journal of Water Resources Planning and Management*. 128(1):12-20.

Chowdhury JU, Stedinger JR, Lu LH. 1991. Goodness-of-Fit Tests for Regional Generalized Extreme Value Flood Distributions. *Water Resources Research*. 27(7):1765-1776.

Cianfrani CM, Hession WC, Rizzo DM. 2006. Watershed imperviousness impacts on stream channel condition in southeastern Pennsylvania. *Journal of the American Water Resources Association*. 42(4):941-956.

Corbane C, Raclot D, Jacob F, Albergel J, Andrieux P. 2008. Remote sensing of soil surface characteristics from a multiscale classification approach. *Catena*. 75(3):308-318.

Dalrymple, T. 1960. Flood-frequency analyses, *Water Supply Paper 1543-A*, U. S. Geological Survey, Washington, D.C.

De Michele C, Rosso R. 2001. Uncertainty assessment of regionalized flood frequency estimates. *Journal of Hydrologic Engineering*. 6(6):453-459.

Detenbeck NE, Brady VJ, Taylor DL, Snarski VM, Batterman SL. 2005. Relationship of stream flow regime in the western Lake Superior basin to watershed type characteristics. *Journal of Hydrology*. 309(1-4):258-276.

Dinpashoh Y, Fakheri-Fard A, Moghaddam M, Jahanbakhsh S, Mirnia M. 2004. Selection of variables for the purpose of regionalization of Iran's precipitation climate using multivariate methods. *Journal of Hydrology*. 297(1-4):109-123.

Duffy CJ. 2004. Semi-discrete dynamical model for mountain-front recharge and water balance estimation: Rio Grande of southern Colorado and New Mexico, in Groundwater Recharge in a Desert Environment: The Southwestern United States, Water Sci. Appl. Ser., vol. 9, JF Hogan, F Phillips, and B Scanlon (eds.), pp. 255 – 271, AGU, Washington, D. C.

Duveiller G, Defourny P, Desclee B, Mayaux P. 2008. Deforestation in Central Africa: Estimates at regional, national and landscape levels by advanced processing of systematically-distributed Landsat extracts. *Remote Sensing of Environment*. 112(5):1969-1981.

Eng K, Tasker GD, Milly PCD. 2005. An analysis of region-of-influence methods for flood regionalization in the gulf-atlantic rolling plains. *Journal of the American Water Resources Association*. 41(1):135-143.

Eng K, Stedinger JR, Gruber AM. 2007a. Regionalization of streamflow characteristics for the Gulf-Atlantic rolling plains using leverage-guided region-of-influence regression, World Environmental & Water Resources Conference - Restoring our Natural Habitat, K.C. Kabbes editor, Tampa, Florida, May 15-18, 2007.

Eng K, Milly PCD, Tasker GD. 2007b. Flood regionalization: A hybrid geographic and predictor-variable region-of-influence regression method, *Journal of Hydrologic Engineering*. **12**(6), 585-591.

Erdas 2010. Erdas Field Guide, Atlanta: Leica Geosystems, 812 p.

Feaster TD, Gotvald AJ, Weaver JC. 2009. Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 3, South Carolina: U.S. Geological Survey *Scientific Investigations Report 2009-5156*, 226 p.

Fill HD, Stedinger JR. 1998. Using regional regression within index flood procedures and an empirical Bayesian estimator. *Journal of Hydrology*. 210(1-4):128-145.

Fovell RG, Fovell MYC. 1993. Climate Zones of the Conterminous United-States Defined Using Cluster-Analysis. *Journal of Climate*. 6(11):2103-2135.

Friendly M, Fox J. 2010. candisc: Generalized Canonical Discriminant Analysis. R package version 0.5-19. http://CRAN.R-project.org/package=candisc

González I, Déjean S. 2009. CCA: Canonical correlation analysis. R package version 1.2. http://CRAN.R-project.org/package=CCA

Gotvald AJ, Feaster TD, Weaver JC. 2009: Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 1, Georgia: U.S. Geological Survey *Scientific Investigations Report 2009–5043*, 120 p.

Griffis VW, Stedinger JR. 2007a. The use of GLS regression in regional hydrologic analyses. *Journal of Hydrology*. 344(1-2):82-95.

Griffis VW, Stedinger JR. 2007b. Evolution of flood frequency analysis with Bulletin 17. *Journal of Hydrologic Engineering*. 12(3):283-297.

Griffis VW, Stedinger JR. 2007c. Log-Pearson Type 3 distribution and its application in flood frequency analysis. I: Distribution characteristics. *Journal of Hydrologic Engineering*. 12(5):482-491.

Griffis VW, Stedinger JR. 2009. Log-Pearson Type 3 Distribution and Its Application in Flood Frequency Analysis. III: Sample Skew and Weighted Skew Estimators. *Journal of Hydrologic Engineering*. 14(2):121-130.

Groupe de Recherche en Hydrologie Statistique (GREHYS). 1996. Presentation and review of some methods for regional flood frequency analysis. *Journal of Hydrology*. 186(1-4):63-84.

Gruber AM, Stedinger JR. 2008. Models of LP3 regional skew, data selection, and Bayesian GLS regression. In proc., World Environmental & Water Resources Conf. 2008, R Babcock and RWalton, eds., ASCE, Reston, Va., Paper No. 596

Gumbel, E.J. 1958 *Statistics of Extremes*, Columbia University Press, New York.

Guse B, Thieken AH, Castellarin A, Merz B. 2010. Deriving probabilistic regional envelope curves with two pooling methods. *Journal of Hydrology*. 380(1-2):14-26.

Hosking JRM, Wallis JR. 1993. Some Statistics Useful in Regional Frequency-Analysis. *Water Resources Research*. 29(2):271-281.

Hosking JRM, Wallis JR. 1988. The Effect of Intersite Dependence on Regional Flood Frequency-Analysis. *Water Resources Research*. 24(4):588-600.

Hosking JRM, Wallis JR. 1997. *Regional Frequency Analysis: An Approach Based on L-Moments*, Cambridge University Press, New York, New York, 224 pp.

Hosking JRM, Wallis JR, Wood EF. 1985. An Appraisal of the Regional Flood Frequency Procedure in the UK Flood Studies Report. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*. 30(1):85-102.

Hosking JRM, Wallis JR, Wood EF. 1985. Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments. *Technometrics*. 27(3):251-261.

Hutchinson KJ, Haynes DA, Schnoor JL. 2010. Human-Impacted Water Resources: Domain Stratification and Mapping to Determine Hydrologically Similar Units. *Environmental Science & Technology*. 44(20):7890-7896.

Ilorme F, Griffis VW, Watkins DW. 2011 Regional Rainfall Frequency Analysis for Flood Risk Assessment in Haiti, *Journal of Natural Hazards*, submitted paper.

Inbar M, Gilichinsky M, Melekestsev I, Melnikov D, Zaretskaya N. 2011. Morphometric and morphological development of Holocene cinder cones: A field and remote sensing study in the Tolbachik volcanic field, Kamchatka. *Journal of Volcanology and Geothermal Research*. 201(1-4):301-311.

Interagency Committee on Water Data (IACWD), Guidelines for Determining Flood Flow Frequency, Bulletin 17B (revised and corrected), 28 pp., Hydrol. Subcomm., Washington, D.C., March 1982.

Jenkinson AF. 1969. Statistics of extremes, in *Estimation of Maximum Floods*, *WMO 233*, *TP 126*, *Tech. Note 98*, chap. 5, pp. 183–228, World Meteorol. Off., Geneva, Switzerland.

Jensen JR. 2005. Introductory digital image processing: a remote sensing perspective, Prentice Hall, Upper Saddle River, N.J.

Jeong D, Stedinger JR, Kim Y, Sung JH. 2007. Bayesian GLS for Regionalization of Flood Characteristics in Korea 2007, World Environmental and Water Resources Congress.

Jin MH, Stedinger JR. 1989. Flood Frequency-Analysis with Regional and Historical Information. *Water Resources Research*. 25(5):925-936.

Johnson RA, Wichern DW. 2007. *Applied Multivariate Statistical Analysis*. Pearson Education, Upper Saddle River, xviii, 773 p. pp.

Jolliffe IT. 2002. *Principal Component Analysis*. Springer Series in Statistics. Springer, New York, xxix, 487 p. pp.

Kahya E, Kalayci S, Piechota TC. 2008. Streamflow regionalization: Case study of Turkey. *Journal of Hydrologic Engineering*. 13(4): 205-214.

Kirby W. 1974. Algebraic Boundedness of Sample Statistics. *Water Resources Research*. 10(2):220-222.

Kjeldsen TR, Rosbjerg D. 2002. Comparison of regional index flood estimation procedures based on the extreme value type I distribution. *Stochastic Environmental Research and Risk Assessment*. 16(5):358-373.

Kjeldsen TR, Jones DA. 2006. Prediction uncertainty in a median-based index flood method using L moments. *Water Resources Research*. 42(7).

Kottegoda KT, Rosso R. 1997. *Statistics, Probability and Reliability for Civil and Environmental Engineers*, McGraw-Hill, New York.

Kroll CN, Stedinger JR. 1998. Regional hydrologic analysis: Ordinary and generalized least squares revisited. *Water Resources Research*. 34(1):121-128.

Kundzewicz ZW, Kaczmarek Z. 2000. Coping with hydrological extremes. *Water International*. 25(1):66-75.

Landwehr JM, Matalas NC, Wallis JR. 1979. Probability Weighted Moments Compared with Some Traditional Techniques in Estimating Gumbel Parameters and Quantiles. *Water Resources Research*. 15(5):1055-1064.

Lettenmaier DP, Wallis JR, Wood EF. 1987. Effect of Regional Heterogeneity on Flood Frequency Estimation. *Water Resources Research*. 23(2):313-323.

Lins HF. 1985. Interannual Streamflow Variability in the United-States Based on Principal Components. *Water Resources Research*. 21(5):691-701.

Lins HF. 1997. Regional streamflow regimes and hydroclimatology of the United States. *Water Resources Research*. 33(7):1655-1667.

MacKinnon D, Tetzlaff D. 2009. Conceptualising Scale in Regional Studies and Catchment Science – Towards an Integrated Characterisation of Spatial Units. *Geography Compass*. 3(3): 976-996.

Madsen H, Rosbjerg D. 1997. Generalized least squares and empirical Bayes estimation in regional partial duration series index-flood modeling. *Water Resources Research*. 33(4):771-781.

Malekinezhad H, Nachtnebel HP, Klik A. 2011. Comparing the index-flood and multiple-regression methods using L-moments. *Physics and Chemistry of the Earth*. 36(1-4):54-60.

Mallants D, Feyen J. 1990. Defining Homogeneous Precipitation Regions by Means of Principal Components-Analysis. *Journal of Applied Meteorology*. 29(9):892-901.

Martins ES, Stedinger JR. 2000. Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*. 36(3):737-744.

Martins ES, Stedinger JR. 2002. Cross correlations among estimators of shape. *Water Resources Research*. 38(11).

Mazvimavi D, Burgers SLGE, Stein A. 2006. Identification of basin characteristics influencing spatial variation of river flows. *International Journal of Applied Earth Observation and Geoinformation*. 8(3):165-172.

McDonnell JJ, Sivapalan M, Vache K, Dunn S, Grant G, Haggerty R, Hinz C, Hooper R, Kirchner J, Roderick ML et al. 2007. Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology. *Water Resources Research*. 43(7).

Mendas A. 2010. The contribution of the digital elevation models and geographic information systems in a watershed hydrologic research. *Applied Geomatics*. Vol 2, pp33-42

Mishra AK, Coulibably P. 2009. Development in hydrometric networks design: A review. *Rev. Geophys*. 47, RG2001.

Moglen GE, Hartman GL. 2001. Resolution effects on hydrologic modeling parameters and peak discharge. *Journal of Hydrologic Engineering*. 6(6):490-497.

National Research Council (NRC). 1988. *Estimating Probabilities of Extreme Floods*, National Academy Press, Washington, D.C., 141 pp.

Natural Environment Research Council (NERC). 1975. *Flood Studies Report*, vol. 1, London.

Neykov NM, Neytchev PN, Van Gelder PHAJM, Todorov VK. 2007. Robust detection of discordant sites in regional frequency analysis. *Water Resources Research*. 43(6).

Panday S, Huyakorn PS. 2004. A fully coupled physically-based spatially-distributed model for evaluating surface/subsurface flow. *Advances in Water Resources*. 27(4):361-382.

Nelson EJ, Paudel M, Scharffenberg W. 2009. Comparison of Lumped and Quasi-Distributed Clark Runoff Models Using the SCS Curve Number Equation. *Journal of Hydrologic Engineering*. 14(10):1098-1106.

Potter KW, Lettenmaier DP. 1990. A Comparison of Regional Flood Frequency Estimation Methods Using a Resampling Method. *Water Resources Research*. 26(3):415-424.

Pryde et al. 2007 Comparison of Watershed boundaries derived from SRTM and ASTER digital elevation datasets and from a digitized topographic map, ASABE Annual International Meeting, Minneapolis, MN.

Rao AR, Srinivas VV. 2006. Regionalization of watersheds by hybrid-cluster analysis. *Journal of Hydrology*. 318(1-4):37-56.

Tang Y, Reed P, van Werkhoven K, Wagener T. 2007. Advancing the identification and evaluation of distributed rainfall-runoff models using global sensitivity analysis. *Water Resources Research*. 43(6).

Reis DS. Jr. 2005. Flood frequency analysis employing Bayesian regional regression and imperfect historical information, PhD. Dissertation, Cornell University, Ithaca, NY.

Reis DS, Stedinger JR, Martins ES. 2005. Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation. *Water Resources Research*. 41(10).

Ries KG, Crouse MY. 2002. The National Flood Frequency Program, Version 3: A computer program for estimating magnitude and frequency of floods for ungaged sites, U.S. Geological Survey *Water Resources Investigations Report 02-4168*, 53 p.

Rhee J, Im J, Carbone GJ, Jensen JR. 2008. Delineation of climate regions using in-situ and remotely-sensed data for the Carolinas. *Remote Sensing of Environment*. 112(6):3099-3111.

Robson A, Reed D. 1999. Flood Estimation Handbook 3: Statistical Procedures of Flood Frequency Estimation, Institute of Hydrology, Wallingford/UK 338 pp.

Saf B. 2009. Regional Flood Frequency Analysis Using L-Moments for the West Mediterranean Region of Turkey. *Water Resources Management*. 23(3):531-551.

Shu C, Ouarda TBMJ. 2008. Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system. *Journal of Hydrology*. 349(1-2):31-43.

Snelder TH, Lamouroux N, Leathwick JR, Pella H, Sauquet E, Shankar U. 2009. Predictive mapping of the natural flow regimes of France. *Journal of Hydrology*. 373(1-2):57-67.

Soil Survey Division Staff. 1993. *Soil Survey Manual*. Soil Conservation Service. U.S. Department of Agriculture Handbook 18.

Srinivas VV, Tripathi S, Rao AR, Govindaraju RS. 2008. Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering. *Journal of Hydrology*. 348(1-2):148-166.

Stedinger JR, Lu LH. 1995. Appraisal of Regional and Index Flood Quantile Estimators. *Stochastic Hydrology and Hydraulics*. 9(1):49-75.

Stedinger JR, Griffis VW. 2008. Flood frequency analysis in the United States: Time to update. *Journal of Hydrologic Engineering*. 13(4):199-204.

Stedinger JR, Vogel RM, Foufoula-Georgiou E. 1993. "Frequency Analysis of Extreme Events", in Handbook of Hydrology, chap. 18, pp. 18.1-18.66, McGraw-Hill Book Co., NY.

Stedinger JR, Tasker GD. 1985. Regional Hydrologic Analysis, 1. Ordinary, Weighted, and Generalized Least-Squares Compared. *Water Resources Research*. 21(9):1421-1432.

Stedinger JR, Tasker GD. 1986a. Correction to "Regional hydrologic analysis 1. Ordinary, weighted, and generalized least squares compared." *Water Resources Research*. 22(5), 844.

Stedinger JR, Tasker GD. 1986b. Regional Hydrologic Analysis, 2, Model-Error Estimators, Estimation of Sigma and Log-Pearson Type-3 Distributions. *Water Resources Research*. 22(10):1487-1499.

Tasker GD, Stedinger JR. 1989. An Operational GLS Model for Hydrologic Regression. *Journal of Hydrology*. 111(1-4):361-375.

Tasker GD, Hodge SA, Barks CS. 1996. Region of influence regression for estimating the 50-year flood at ungaged sites. *Water Resources Bulletin*. 32(1):163-170.

Tasker GD. 1980. Hydrologic Regression with Weighted Least-Squares. *Water Resources Research*. 16(6):1107-1113.

Tetzlaff D, McDonnell JJ, Uhlenbrook S, McGuire KJ, Bogaart PW, Naef F, Baird AJ, Dunn SM, Soulsby C. 2008. Conceptualizing catchment processes: simply too complex? *Hydrological Processes*. 22(11):1727-1730.

Todorov V, Filzmoser P. 2009. An Object-Oriented Framework for Robust Multivariate Analysis. *Journal of Statistical Software*. 32(3):1-47.

Tsakiris G, Nalbantis I, Cavadias G. 2011 Regionalization of low flows based on Canonical Correlation Analysis, *Advances in Water Resources*. 34(7): 865-872, ISSN 0309-1708, DOI: 10.1016/j.advwatres.2011.04.007.

U.S. Army Corps of Engineers (USACE). 1994. "Engineering and Design – Hydrologic Engineering Analysis Concepts for Cost-Shared Flood Damage Reduction Studies," Engineer Pamphlets EP 1110-2-10. Proponent: CECW-EH-Y.

U.S. Department of Agriculture, Natural Resources Conservation Service. *National Soil Survey Handbook*, title 430-VI. Available online at http://soils.usda.gov/technical/handbook/. Accessed [07/18/2011].

US. Geological Survey (USGS). 1995. State Soil Geographic (STATSGO) data base for the conterminous United States http://www.data.gov/geodata/g601440 . Accessed [07/08/2011].

Venables WN, Ripley BD. 2002. *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Viglione A, Laio F, Claps P. 2007. A comparison of homogeneity tests for regional frequency analysis. *Water Resources Research*. 43(3).

Viglione A. 2010. nsRFA: Non-supervised Regional Frequency Analysis. R package Version 0.7-0. http://CRAN.R-project.org/package=nsRFA

Vogel RM, Wilson I. 1996. Probability distribution of annual maximum, mean, and minimum streamflows in the United States, *Journal of Hydrologic Engineering*. 1(2), 69-76.

Wallis JR, Wood EF. 1985. Relative Accuracy of Log Pearson-III Procedures. *Journal of Hydraulic Engineering*. 111(7):1043-1056.

Wallis JR. 1988. Catastrophes, Computing and Containment: Living in our restless habitat, *Speculation in Science and Technology. 11*(4), 295-315.

Wallis JR, Matalas NC, Slack JR. 1974. Just a Moment! *Water Resources Research*. 10(2), 211-221.

Weaver JC, Feaster TD, Gotvald AJ. 2009. Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 2, North Carolina: U.S. Geological Survey *Scientific Investigations Report 2009–5158*, 113 p.

Yeung KY, Ruzzo WL. 2001. Principal component analysis for clustering gene expression data. *Bioinformatics*. 17(9):763-774.

Zrinji Z, Burn DH. 1994. Flood Frequency-Analysis for Ungauged Sites Using a Region of Influence Approach. *Journal of Hydrology*. 153(1-4):1-21.

Appendix A

Soil Drainage and Infiltration Indices

This appendix describes the soil drainage (SI) and infiltration (Inf) indices employed as watershed attributes in the regional flood frequency techniques employed in this dissertation.

A1. Soil Drainage Index

The **soil drainage index** is a numerical representation of the drainage class which identifies the natural drainage properties of the soil. In particular, it refers to the frequency and duration of wet periods under conditions similar to those under which the soil formed (*National Soil Survey Handbook*, accessed 07/18/2011). There are seven major classes of soil drainage as described in the *Soil Survey Manual* (1993):

Excessively drained (E): Water is very rapidly removed from the soil. Soils are generally very coarse textured, rocky, or shallow. Some are steep. All are free of the mottling related to wetness.

Somewhat excessively drained (SE): Water is rapidly removed from the soil. Soils are generally sandy and rapidly pervious. Some are shallow. Some are steep enough that much of the water they receive is lost as runoff. All are free of the mottling related to wetness.

Well drained (W): Water is readily removed from the soil, but not rapidly. Water is available to plants throughout most of the growing season, and wetness does not inhibit growth of roots for significant periods during most growing seasons. Well drained soils are generally medium textured and are mainly free of mottling.

Moderately well drained (MW): Water is removed from the soil somewhat slowly during some periods. Soils are wet for only a short time during the growing season, but periodically they are wet long enough that most mesophytic crops are affected. They typically have a slowly pervious layer

within or directly below the solum, or periodically receive high rainfall, or both.

Somewhat poorly drained (SP): Water is removed slowly enough that the soil is wet for significant periods during the growing season. Wetness markedly restricts the growth of mesophytic crops unless artificial drainage is provided. Soils typically have a slowly pervious layer, a high water table, additional water from seepage, nearly continuous rainfall, or a combination of these.

Poorly drained (P): Water is removed so slowly that the soil is saturated periodically during the growing season or remains wet for long periods. Free water is commonly at or near the surface for long enough during the growing season that most mesophytic crops cannot be grown unless the soil is artificially drained. The soil is not continuously saturated in layers directly below plow depth. Poor drainage results from a high water table, a slowly pervious layer within the profile, seepage, nearly continuous rainfall, or a combination of these.

Very poorly drained (VP): Water is removed from the soil so slowly that free water remains at or on the surface during most of the growing season. Unless the soil is artificially drained, most mesophytic crops cannot be grown. Soils are commonly level or depressed and are frequently ponded. Yet, where rainfall is high and nearly continuous, they can have moderate or high slope gradients.

Herein, numerical representations of these drainage classes are necessary for the application of multivariate statistics and regional regression techniques. The values of the soil drainage index (SI) and the corresponding drainage class are as follows: 1 represents class E, 2 represents class SE, 3 represents class W, 4 represents class MW, 5 represents class SP, 6 represents class P, and 7 represents class VP.

A2. Infiltration Index

The **infiltration index** is a numerical representation of the hydrologic soil group which identifies the runoff potential of the soil and the minimum infiltration rate following prolonged periods of wetness (*National Soil Survey Handbook*, accessed

07/18/2011). There are four major groups (A, B, C, and D) defined for classification of soils within the United States. The properties of these groups are defined in the *National Engineering Handbook* (available online at http://www.mi.nrcs.usda.gov/technical/engineering/neh.html, accessed 07/19/2011):

> **Group A:** Soils with high infiltration rates and low runoff potential when thoroughly wet. These soils have high water transmission rates, consisting primarily of deep, well drained to excessively drained sands or gravelly sands.
>
> **Group B:** Soils with moderate infiltration rates when thoroughly wet. These soils have moderate water transmission rates, consisting primarily of moderately deep or deep, moderately well drained or well drained soils that have moderately fine texture to moderately coarse texture.
>
> **Group C:** Soils with slow infiltration rates when thoroughly wet. These soils have slow water transmission rates, consisting primarily of soils with a layer that impedes the downward movement of water or soils of moderately fine texture or fine texture.
>
> **Group D:** Soils with very slow infiltration rates and high runoff potential when thoroughly wet. These soils have very slow water transmission rates, consisting primarily of clays with a high shrink-swell potential, soils with a high water table, soils with a claypan or clay layer at or near the surface, and shallow soils over nearly impervious material.

Herein, numerical representations of these soil groups are necessary for the application of multivariate statistics and regional regression techniques. The values of the infiltration index (Inf) and the corresponding soil group are as follows: 1 represents group A, 2 represents group B, 3 represents group C, and 4 represents group D.

Appendix B

Application of the Physical Discordancy Test

This appendix contains a full description of the procedure to identify physically discordant sites using the metric introduced in Section 3.1.1. A detailed description of the test is provided using an example for Region 2 of Set 1 (N = 22 sites) delineated using the Ward clustering technique applied to mean elevation, basin slope, and soil drainage (see Figure 3.5).

For a given region, the first step is to perform a principal component analysis using the normalized and standardized physical characteristics for each site (or watershed) therein. This is accomplished using equations (32), (33), and (34) to compute the principal component functions (Y_i), and the associated eigenvectors (e_i) and variance (eigenvalues, λ_i). For the purpose of the physical discordancy test, λ_i and Y_i are the only parameters of interest. For Region 2 of Set 1, Table B.1 reports the principal component loadings, the standard deviation of each component ($S_i = \sqrt{\lambda_i}$) and the cumulative variance explained when higher order PCs are retained. Nine principal components are obtained as nine physical characteristics are investigated herein as possible indicators of extreme hydrologic response. Only the components whose standard deviations are greater than 1 are to be employed in the physical discordancy test, as these are the PCs which explain the majority of the variability in the original dataset (Johnson and Wichern 2007, p. 451). For Region 2, only the first four PCs should be used; these PCs explain 81% of the variation of the dataset.

Table B.1
PCA output obtained using normalized and standardized physical variables at sites in Region 2 of Set 1.

Physical Characteristic	**PC_1**	**PC_2**	**PC_3**	**PC_4**	**PC_5**	**PC_6**	**PC_7**	**PC_8**	**PC_9**
Drainage Area	0.51	0.15	-0.01	0.21	0.29	-0.04	0.27	0.72	-0.03
Main Channel Slope	-0.48	-0.26	0.28	0.11	-0.05	-0.27	0.70	0.12	0.20
Basin Slope	-0.05	-0.17	-0.64	-0.58	0.23	0.04	0.34	0.00	-0.22
Basin Shape	0.53	0.20	-0.14	0.05	-0.15	-0.05	0.42	-0.52	0.43
Basin Elevation	-0.06	-0.26	-0.15	0.54	0.70	0.06	-0.02	-0.34	-0.05
% Impervious Cover	-0.13	0.58	-0.10	0.16	0.00	-0.55	0.10	-0.16	-0.51
% Forest Cover	-0.35	0.44	-0.22	-0.12	0.31	-0.14	-0.22	0.16	0.66
Soil Drainage Index	-0.28	0.22	-0.43	0.45	-0.37	0.55	0.18	0.13	-0.05
Infiltration Index	-0.09	0.43	0.47	-0.27	0.34	0.54	0.25	-0.15	-0.15
$S_i = \sqrt{\lambda_i}$	1.65	1.52	1.07	1.05	0.97	0.57	0.44	0.42	0.27
Cumulative Variance:	0.30	0.56	0.69	0.81	0.92	0.95	0.97	0.99	1.00

The first four principal component scores (with eigenvalues greater than 1) are then evaluated for all sites in Region 2 using the loadings in Table B.1 in conjunction with the normalized and standardized values of the physical variables at the site in question. The resulting PC scores and corresponding Station ID are reported in Table B.2. In order to identify extreme values of the PC scores, critical values ($T_{critical}$) are computed using equation (54) wherein $S = \sqrt{\lambda_i}$ for i = 1, …, 4, and $t_{0.975,N-1} = 2.08$ as Region 2 contains N = 22 sites. The critical values for Region 2 are listed in Table B.3.

Table B.2
First four principal component scores for each watershed in Region 2 of Set 1.
Bold font indicates physically discordant watersheds.

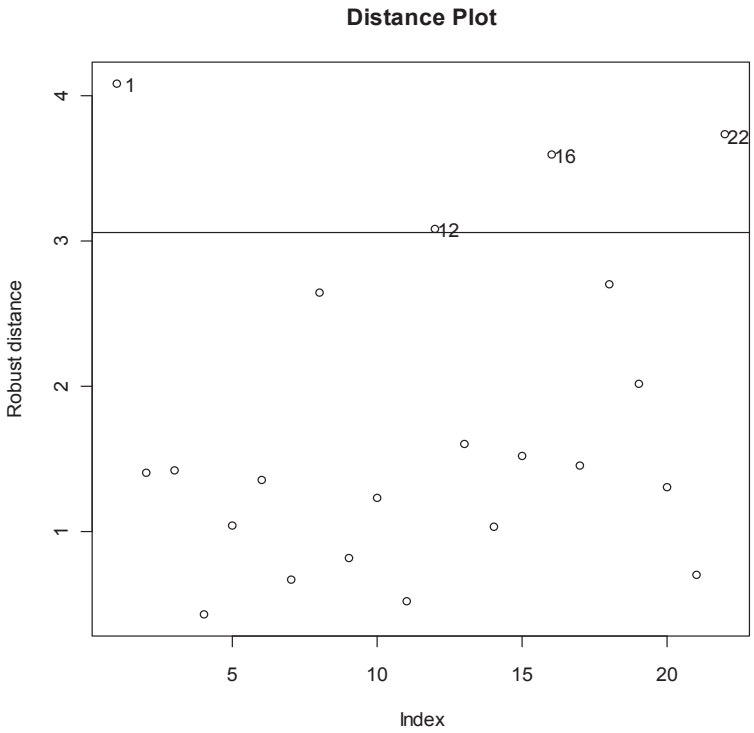
Station ID	PC_1	PC_2	PC_3	PC_4
88	-1.90	0.41	-0.82	0.65
145	1.47	**-3.92**	-0.23	0.78
146	0.41	0.55	-0.62	1.77
147	-0.37	0.29	-1.13	1.17
177	-0.21	0.86	-0.86	1.11
178	-1.61	1.42	0.79	0.72
179	2.85	**-3.44**	0.76	0.07
180	-0.89	1.57	0.32	-0.31
181	1.74	2.28	-0.29	0.77
226	**3.64**	0.47	-1.18	-0.31
227	-1.94	-1.78	-1.93	-1.23
228	0.73	0.69	0.74	**-2.30**
229	-0.15	0.49	-0.49	-2.13
231	-1.12	-0.92	-0.88	-1.15
232	0.87	0.57	1.18	1.21
233	-1.16	-0.36	0.68	-0.26
234	-0.38	-0.53	1.52	-0.16
236	-1.19	-0.94	**2.54**	0.62
240	-1.98	-0.50	-1.47	0.74
246	0.83	2.03	0.68	-0.69
247	2.66	1.29	-0.25	-0.60
248	-2.29	-0.52	0.95	-0.49

Table B.3
Critical values for the first four principal component scores for Region 2 in Set 1.

	PC_1	PC_2	PC_3	PC_4
$T_{critical}$	$\pm$3.64	$\pm$3.16	$\pm$2.23	$\pm$2.18

The values of each PC score in Table B.2 are then compared against the corresponding critical values in Table B.3. Any PC scores which are less than $-T_{critical}$ or greater than $+T_{critical}$ are deemed extreme with regards to the characteristics generally observed in the region, and the corresponding watershed is identified as physically discordant. The discordant sites identified in Region 2 are indicated by bold font in Table B.2.

As the standardized procedure for region delineation outlined in Section 3.1 recommends only removing those sites which are both hydrologically and physically discordant, the results above are then compared with those of the robust hydrologic discordancy test proposed by Neykov et al. (2007). Details of the latter test are provided by Neykov et al. (1997); herein, the test is implemented in R using the package provided by Todorov and Filzmoser (2009). An output of the test is a distance plot which shows the location of the discordant sites using an internal index (Neykov ID = 1, …., N) related to the order in which the sites where entered in the test. The distance plot for Region 2 of Set 1 is illustrated in Figure B.1. Four sites are identified as hydrologically discordant; the Station IDs of these sites are tabulated in Figure B.1. The site(s) identified as both hydrologically and physically discordant (HPD) are to be removed from the region prior to developing the regional flood distribution via the GEV/L-moment Index Flood procedure. For Region 2, only watershed 228 is identified as HPD.



Neykov id	Station id
1	88
12	228
16	233
22	248

Figure B.1: Distance plot obtained from the hydrological discordancy test of Neykov et al. (2007) and Station IDs of sites identified as hydrologically discordant.

Appendix C

Additional Results of Jackknife Resampling Procedure

This appendix presents additional results for the jackknife resampling analyses conducted in Chapters 3 and 4. These results correspond to either the original regions are reported for regions from which all hydrologically discordant (HD) sites have been removed, as well as regions from which all sites which are both hydrologically and physically discordant (HPD) have been removed. Results for regions without HD sites are included herein for comparison purposes only.

Table C.1
MSE of flood quantiles obtained for regions delineated in Case 1 (Chapter 3).

	T = 10 years (p = 0.90)		**T = 25 years (p = 0.96)**		**T = 50 years (p = 0.98)**		**T = 100 years (p = 0.99)**	
	HPD	HD	HPD	HD	HPD	HD	HPD	HD
Region 1	0.003	0.002	0.009	0.005	0.020	0.011	0.037	0.021
Region 2	0.003	0.004	0.004	0.004	0.012	0.012	0.028	0.029
Region 3	0.002	0.001	0.002	0.001	0.008	0.005	0.020	0.012
Average	0.003	0.002	0.005	0.003	0.013	0.010	0.028	0.021

Table C.2
MSE of flood quantiles obtained for regions delineated in Case 2 (Chapter 3).

	T = 10 years (p = 0.90)		**T = 25 years (p = 0.96)**		**T = 50 years (p = 0.98)**		**T = 100 years (p = 0.99)**	
	HPD	HD	HPD	HD	HPD	HD	HPD	HD
Region 1	0.006	0.004	0.018	0.015	0.034	0.028	0.057	0.047
Region 2	0.003	0.002	0.011	0.010	0.027	0.023	0.052	0.041
Region 3	0.003	0.003	0.017	0.017	0.039	0.035	0.072	0.060
Region 4	0.005	0.004	0.025	0.029	0.061	0.065	0.117	0.120
Region 5	0.008	0.008	0.034	0.036	0.074	0.076	0.135	0.135
Average	0.005	0.004	0.021	0.021	0.047	0.045	0.087	0.081

Table C.3
MSE of flood quantiles obtained for regions delineated in Case 3 (Chapter 3).

	T = 10 years (p = 0.90)		**T = 25 years (p = 0.96)**		**T = 50 years (p = 0.98)**		**T = 100 years (p = 0.99)**	
	HPD	HD	HPD	HD	HPD	HD	HPD	HD
Region 1	0.006	0.005	0.019	0.014	0.037	0.027	0.063	0.045
Region 2	0.004	0.003	0.021	0.013	0.045	0.027	0.082	0.050
Region 3	0.005	0.004	0.025	0.029	0.061	0.065	0.117	0.120
Region 4	0.008	0.008	0.035	0.035	0.081	0.081	0.156	0.156
Region 5	0.010	0.008	0.032	0.033	0.063	0.062	0.109	0.101
Average	0.006	0.006	0.026	0.025	0.057	0.052	0.105	0.094

Table C.4
MSE of flood quantiles obtained for regions delineated in Case 4 (Chapter 3).

	T = 10 years (p = 0.90)		**T = 25 years (p = 0.96)**		**T = 50 years (p = 0.98)**		**T = 100 years (p = 0.99)**	
	HPD	HD	HPD	HD	HPD	HD	HPD	HD
Region 1	0.004	0.004	0.012	0.013	0.023	0.024	0.037	0.041
Region 2	0.005	0.004	0.017	0.015	0.031	0.030	0.050	0.050
Region 3	0.006	0.004	0.024	0.013	0.056	0.031	0.106	0.059
Region 4	0.006	0.004	0.022	0.025	0.049	0.055	0.092	0.101
Region 5	0.007	0.007	0.031	0.031	0.067	0.067	0.122	0.122
Average	0.006	0.005	0.021	0.017	0.045	0.041	0.081	0.075

Table C.5
MSE of flood quantiles obtained for regions delineated in Case 5 (Chapter 3).

	T = 10 years (p = 0.90)		**T = 25 years (p = 0.96)**		**T = 50 years (p = 0.98)**		**T = 100 years (p = 0.99)**	
	HPD	HD	HPD	HD	HPD	HD	HPD	HD
Region 1	0.002	0.002	0.009	0.010	0.017	0.020	0.032	0.036
Region 2	0.005	0.003	0.017	0.011	0.030	0.022	0.048	0.036
Region 3	0.006	0.003	0.017	0.017	0.043	0.041	0.089	0.077
Region 4	0.004	0.003	0.013	0.010	0.029	0.024	0.054	0.046
Region 5	0.003	0.003	0.017	0.017	0.039	0.035	0.072	0.060
Region 6	0.007	0.008	0.031	0.035	0.070	0.069	0.132	0.118
Region 7	0.008	0.008	0.034	0.034	0.067	0.067	0.114	0.114
Average	0.005	0.004	0.020	0.019	0.042	0.040	0.077	0.070

Table C.6
MSE of flood quantiles obtained for regions delineated in Set 2 using aggregated values of basin elevation, basin slope, and soil drainage.

	T = 10 years (p = 0.90)		**T = 25 years (p = 0.96)**		**T = 50 years (p = 0.98)**		**T = 100 years (p = 0.99)**	
	HPD	HD	HPD	HD	HPD	HD	HPD	HD
Region 1	0.005	0.002	0.026	0.012	0.059	0.032	0.112	0.066
Region 2	0.007	0.005	0.027	0.030	0.063	0.066	0.121	0.119
Region 3	0.005	0.002	0.019	0.011	0.004	0.025	0.073	0.050
Region 4	0.004	0.004	0.012	0.012	0.023	0.023	0.040	0.040
Region 5	0.006	0.004	0.024	0.012	0.056	0.023	0.106	0.040
Region 6	0.013	0.012	0.038	0.025	0.077	0.042	0.136	0.069
Average	0.007	0.005	0.024	0.017	0.047	0.035	0.098	0.064

Table C.7
MSE of flood quantiles obtained for regions delineated in Set 1 using spatially distributed representations of basin elevation, basin slope, and soil drainage.

	T = 10 years (p = 0.90)		**T = 25 years (p = 0.96)**		**T = 50 years (p = 0.98)**		**T = 100 years (p = 0.99)**	
	HPD	HD	HPD	HD	HPD	HD	HPD	HD
Region 1	0.003	0.003	0.012	0.013	0.024	0.027	0.041	0.046
Region 2	0.005	0.003	0.022	0.011	0.047	0.023	0.086	0.041
Region 3	0.003	0.003	0.017	0.016	0.039	0.034	0.070	0.060
Region 4	0.011	0.009	0.041	0.043	0.089	0.095	0.169	0.176
Region 5	0.006	0.006	0.028	0.025	0.060	0.050	0.101	0.084
Region 6	0.005	0.004	0.018	0.016	0.033	0.032	0.054	0.054
Region 7	0.005	0.002	0.012	0.015	0.031	0.036	0.065	0.071
Average	0.005	0.004	0.021	0.020	0.046	0.042	0.084	0.076

Table C.8
MSE of flood quantiles obtained for regions delineated in Set 2 using spatially distributed representations of basin elevation, basin slope, and soil drainage.

	T = 10 years (p = 0.90)		**T = 25 years (p = 0.96)**		**T = 50 years (p = 0.98)**		**T = 100 years (p = 0.99)**	
	HPD	HD	HPD	HD	HPD	HD	HPD	HD
Region 1	0.007	0.005	0.038	0.025	0.087	0.054	0.169	0.099
Region 2	0.007	0.005	0.027	0.030	0.063	0.066	0.121	0.119
Region 3	0.004	0.004	0.015	0.015	0.034	0.028	0.065	0.045
Region 4	0.004	0.003	0.015	0.014	0.031	0.030	0.057	0.055
Region 5	0.003	0.002	0.020	0.021	0.055	0.059	0.110	0.117
Region 6	0.010	0.001	0.028	0.010	0.055	0.025	0.093	0.049
Average	0.006	0.003	0.024	0.019	0.054	0.044	0.103	0.081

Appendix D

Additional Information for Regression Model Development

This appendix presents additional statistics related to the ordinary least squares (OLS) regression models of the coefficient of variation, smoothing functions to describe the correlation-distance relationship, and generalized least squares (GLS) regression models of the mean employed in Chapter 5. The statistics are summarized for each of the seven regions delineated for Set 1. Residuals plots showing the adequacy of the regression models are also provided.

D1. OLS Regression Models of Coefficient of Variation

The OLS regression models of the coefficient of variation presented in Table 5.1 were obtained using the function "lm" in R. For each region, the base-10 logarithms of the coefficient of variation and the base-10 logarithms of the area (see equation 60) computed for each site served as inputs to the function. Table D.1 reports the values of the model coefficients derived within each region, as well as the standard error associated with each coefficient, and t-values and p-values resulting from a two-sided hypothesis test to indicate their significance. These results suggest that the coefficient in most regions would be adequately described by the mean regional value; the coefficient on the area is only significant at the 5% level in Region 6. Table D.2 reports summary statistics related to the precision of the models derived for each region. Overall, none of the

models are adequate for prediction purposes; however, the models are sufficient for the purpose of the estimator of coefficient of variation used herein.

Table D.1

Summary statistics for OLS regression model coefficients for each region in Set 1.

	Coefficient	Standard Error	T-value	P-value
		Region 1		
Constant	-0.9391	0.0993	-9.456	0.000
Area	0.0547	0.0473	1.157	0.274
		Region 2		
Constant	-0.9128	0.1005	-9.079	0.000
Area	0.0109	0.0109	0.047	0.871
		Region 3		
Constant	-0.5755	0.1462	-3.938	0.001
Area	-0.1078	0.0748	-1.441	0.1652
		Region 4		
Constant	-0.7600	0.0740	-10.27	0.000
Area	-0.0600	-0.0600	0.036	0.113
		Region 5		
Constant	-0.8480	0.2407	-3.524	0.002
Area	0.0027	0.1000	0.027	0.979
		Region 6		
Constant	-0.6244	0.0998	-6.256	0.000
Area	-0.0934	0.0455	0.046	0.050
		Region 7		
Constant	-0.6430	0.1383	-4.649	0.001
Area	-0.0237	-0.0237	0.0603	0.701

Table D.2
Summary statistics for the OLS regression models of the coefficient of variation derived for each region in Set 1.

	R^2	Adjusted-R^2	Residual Standard Error
Region 1	0.1180	0.0298	0.0709
Region 2	0.0003	-0.0471	0.1040
Region 3	0.0940	0.0487	0.1174
Region 4	0.1014	0.0640	0.0924
Region 5	0.0000	-0.0555	0.1334
Region 6	0.1442	0.1099	0.1163
Region 7	0.0127	-0.0696	0.1214

D2. Smoothing Functions for Estimation of Cross-Correlation

To avoid correlation between the residuals, the cross-correlations used to estimate the sampling covariance matrix are computed using a smoothing function which approximates the correlation between two sites as a function of the distance between the sites. The form of the smoothing function is provided in equation (59). The function parameters (α and θ) were estimated using a non-linear regression method in R via the function "nls". Figure 5.1 illustrates the smoothing function derived for Region 1 of Set 1. The smoothing functions for the other six regions are illustrated in the figures below.

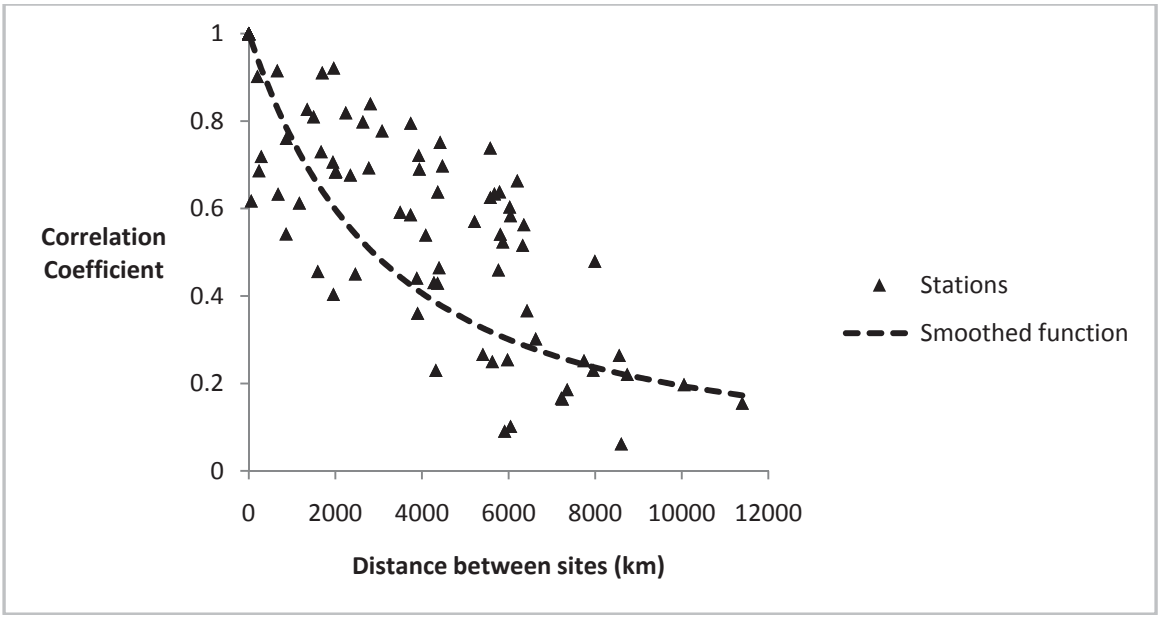


Figure D.1: Correlation-distance smoothing function for Region 2 (α = 0.000083 and θ = 0.9997).

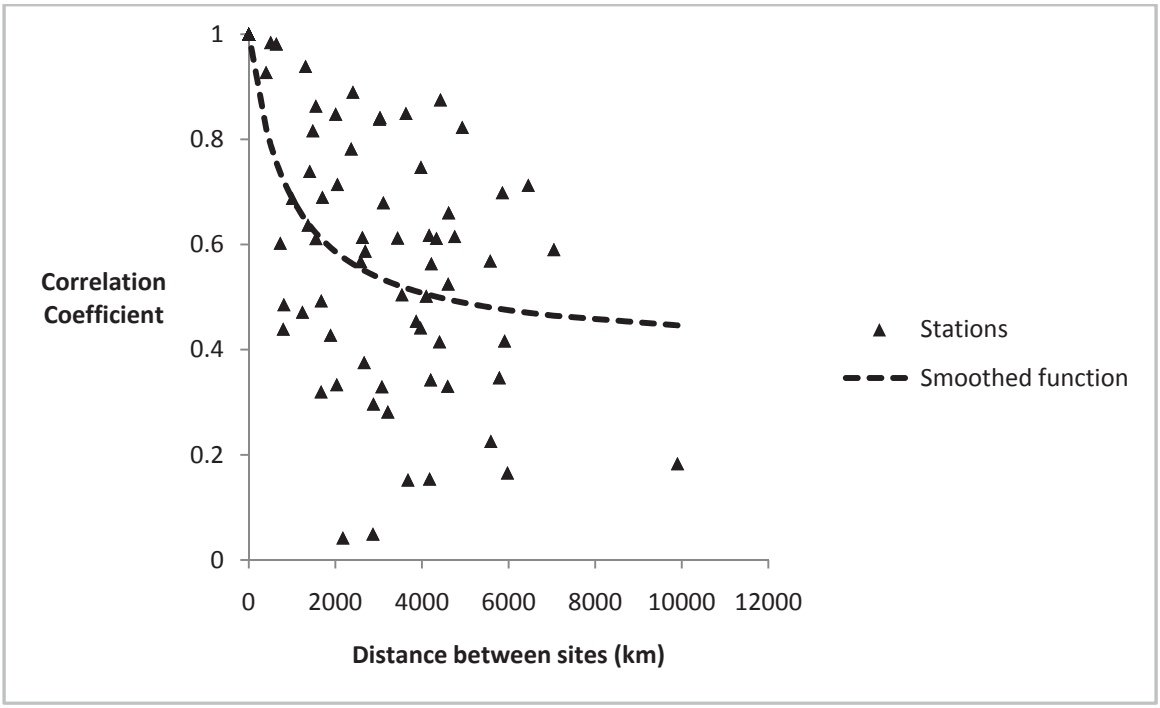


Figure D.2: Correlation-distance smoothing function for Region 3 (α = 0.000679 and θ = 0.9994)

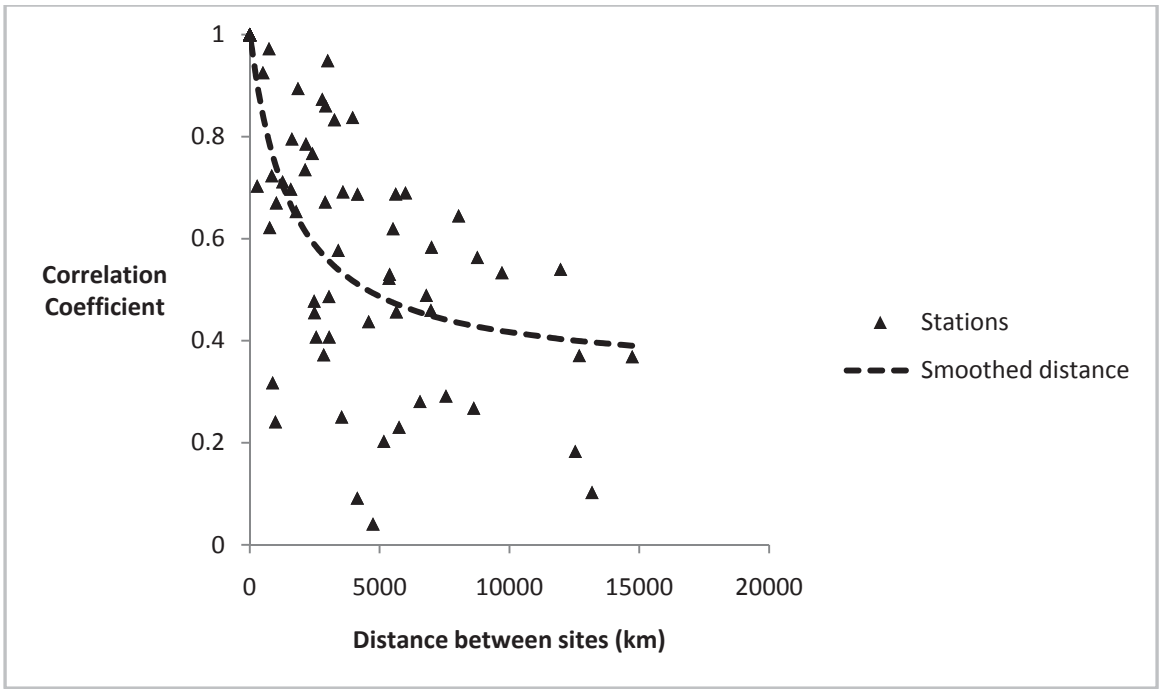


Figure D.3: Correlation-distance smoothing function for Region 4 (α = 0.0003623 and θ = 0.9996).

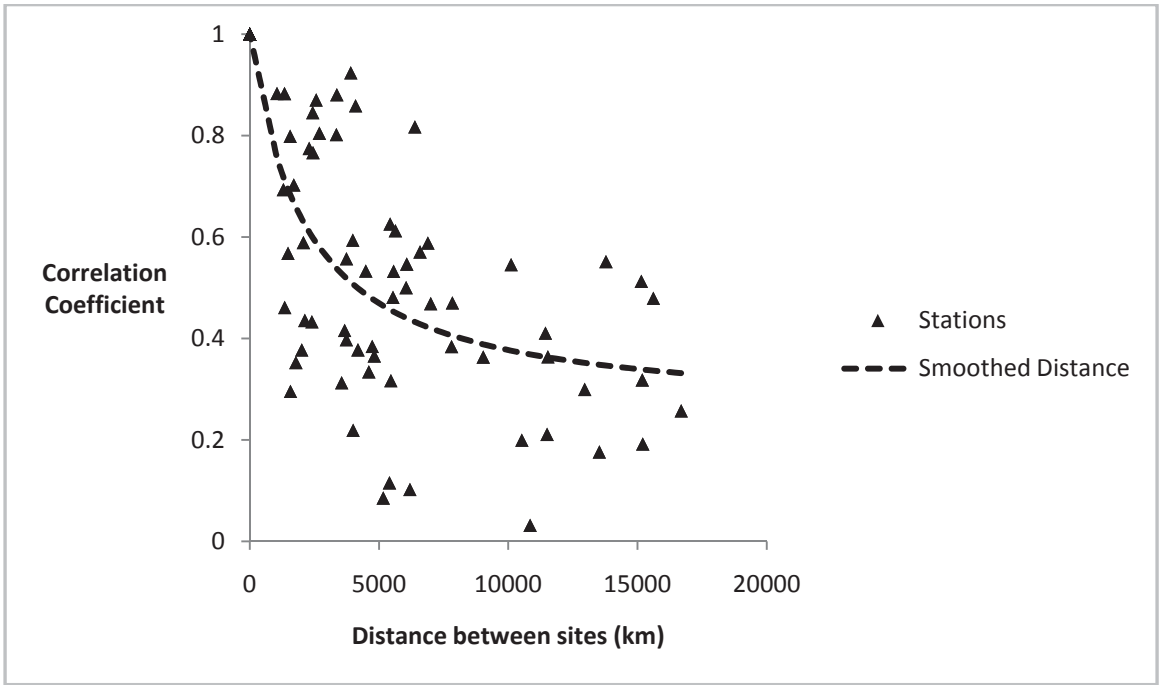


Figure D.4: Correlation-distance smoothing function for Region 5 (α = 0.000245 and θ = 0.9997).

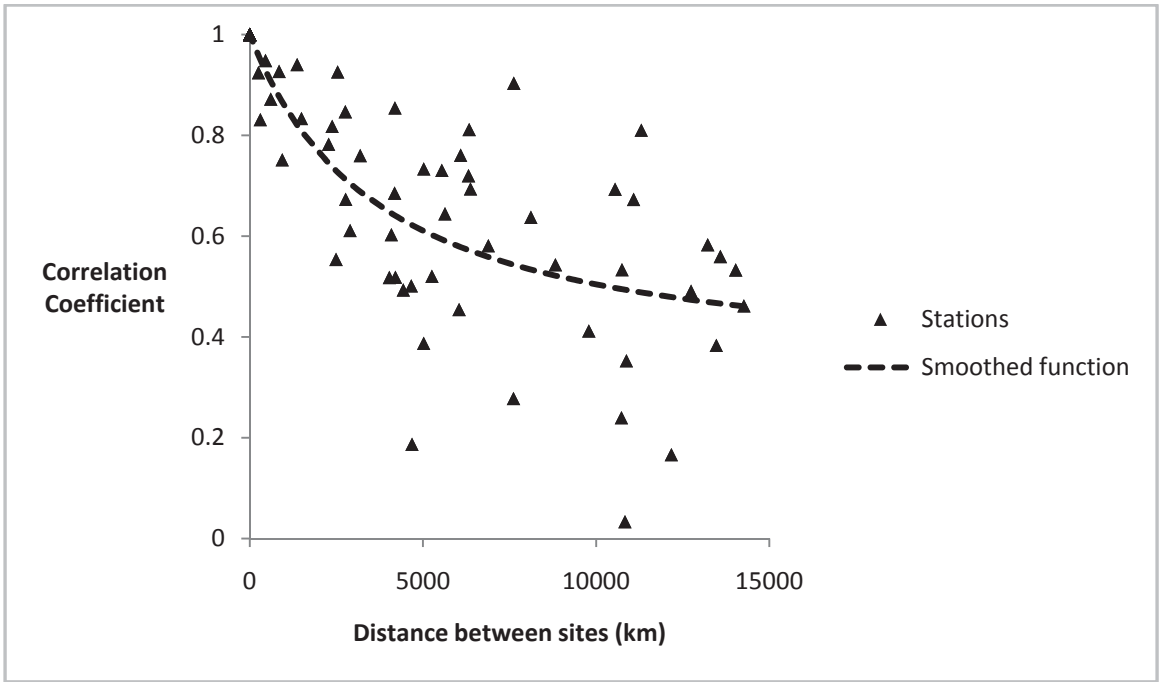


Figure D.5: Correlation-distance smoothing function for Region 6 (α = 0.000157 and θ = 0.9998).

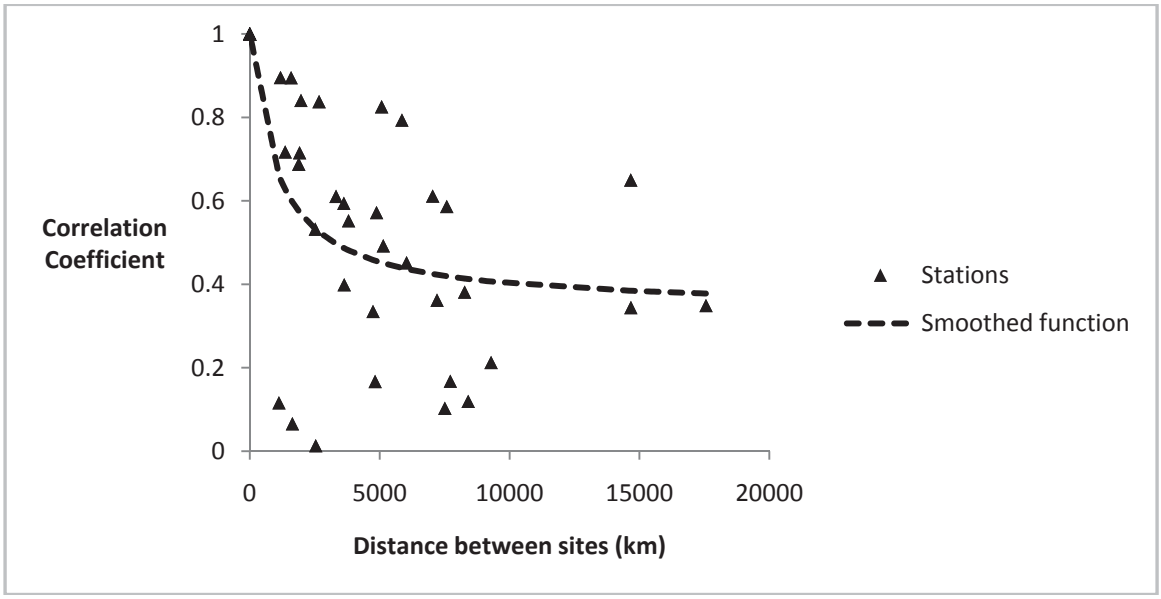


Figure D.6: Correlation-distance smoothing function for Region 7 (α = 0.000565 and θ = 0.9994).

D3. GLS Regression Models of the Mean

The GLS regression models of the mean presented in Table 5.2 were obtained using a modified version of a MATLAB code prepared by Griffis and Stedinger (2007[a]). The tables below report the values of the model coefficients derived for each region, as well as the standard error associated with each coefficient, and t-values and p-values resulting from a two-sided hypothesis test to indicate their significance. Plots of the residuals resulting from the GLS regression models are also provided herein. With the exception of Region 3 wherein the model for the mean was relatively poor, the residuals are all centered on a mean of zero and no patterns are evident; therefore, the regression functions obtained from the GLS model are appropriate.

Table D.3
Summary statistics for GLS regression model coefficients in Region 1.

	Model Coefficient	Standard Error	T-value	P-value
Constant	-3.610	0.9446	-3.820	0.0051
Drainage Area	0.8224	0.0358	22.99	0.0000
% Forest Cover	2.6441	0.4598	5.751	0.0004
Infiltration	1.336	0.2210	56.05	0.0003

Table D.4
Summary statistics for GLS regression model coefficients in Region2.

	Model Coefficient	Standard Error	T-value	P-value
Constant	6.150	1.134	5.421	0.0000
Drainage Area	0.7440	0.1077	6.909	0.0000
Basin Shape	-0.4659	0.1912	-2.436	0.0255
Soil Drainage	-7.260	2.244	-3.236	0.0046

Table D.5
Summary statistics for GLS regression model coefficients in Region 3.

	Model Coefficient	Standard Error	T-value	P-value
Constant	0.8986	0.6980	1.287	0.2152
Drainage Area	0.9538	0.2262	4.216	0.0006
Channel Slope	0.7103	0.3248	2.187	0.0430
% Impervious	0.1236	0.0311	3.976	0.0010
% Forest Cover	-0.0732	0.0330	-2.221	0.0402

Table D.6
Summary statistics for GLS regression model coefficients in Region 4.

	Model Coefficient	Standard Error	T-value	P-value
Constant	2.061	0.2277	9.052	0.0000
Drainage Area	0.6910	0.0458	15.10	0.0000
Basin Shape	-0.3377	0.0926	-3.646	0.0014
Infiltration	1.035	0.0431	2.400	0.0253

Table D.7
Summary statistics for GLS regression model coefficients in Region 5.

	Model Coefficient	Standard Error	T-value	P-value
Constant	-4.386	1.758	-2.495	0.0239
Drainage Area	0.5682	0.0744	7.642	0.0000
Precipitation	2.289	0.9797	2.336	0.0328
% Forest Cover	1.669	0.3186	5.240	0.0001

Table D.8
Summary statistics for GLS regression model coefficients in Region 6.

	Model Coefficient	Standard Error	T-value	P-value
Constant	-8.734	2.018	-4.329	0.0003
Drainage Area	0.5861	0.0495	11.83	0.0000
Basin Slope	0.5916	0.1825	3.241	0.0038
Precipitation	5.772	1.202	4.804	0.0001
Infiltration	2.061	0.6081	3.389	0.0026

Table D.9
Summary statistics for GLS regression model coefficients in Region 7.

	Model Coefficient	Standard Error	T-value	P-value
Constant	5.840	0.7257	8.046	0.0000
Drainage Area	0.4007	0.0782	5.121	0.0006
Basin Shape	-0.8730	0.3207	-2.723	0.0235
Soil Drainage	-10.52	3.129	-3.363	0.0084
Infiltration	9.212	3.294	2.796	0.0208

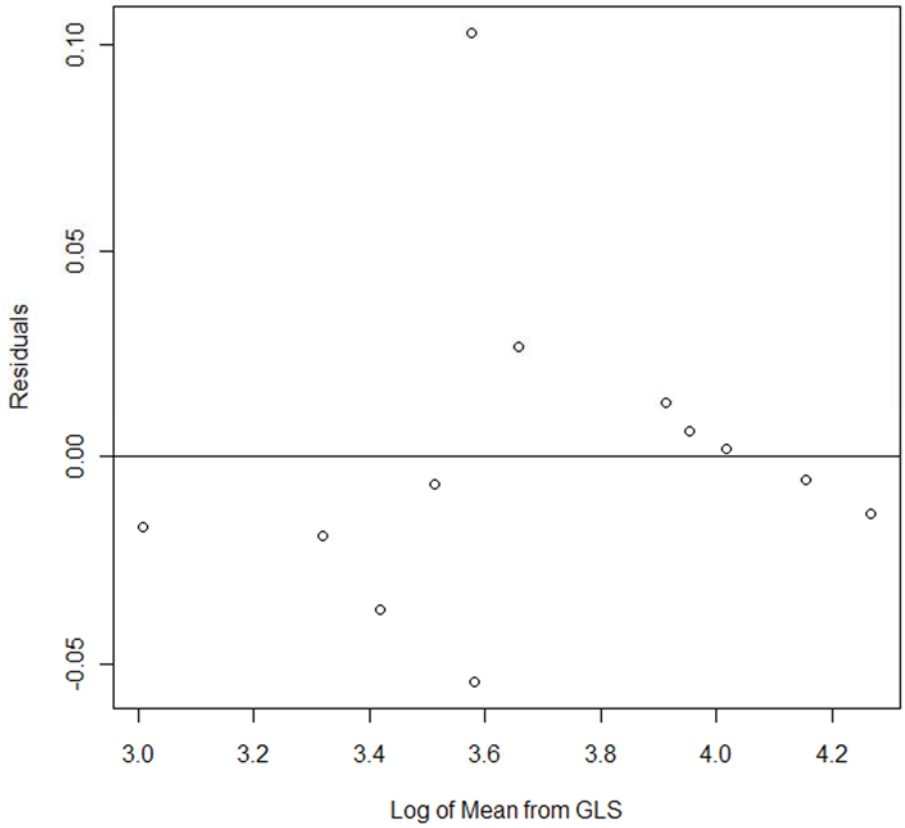


Figure D.7: Plot of residuals for GLS regression model of the mean for Region 1.

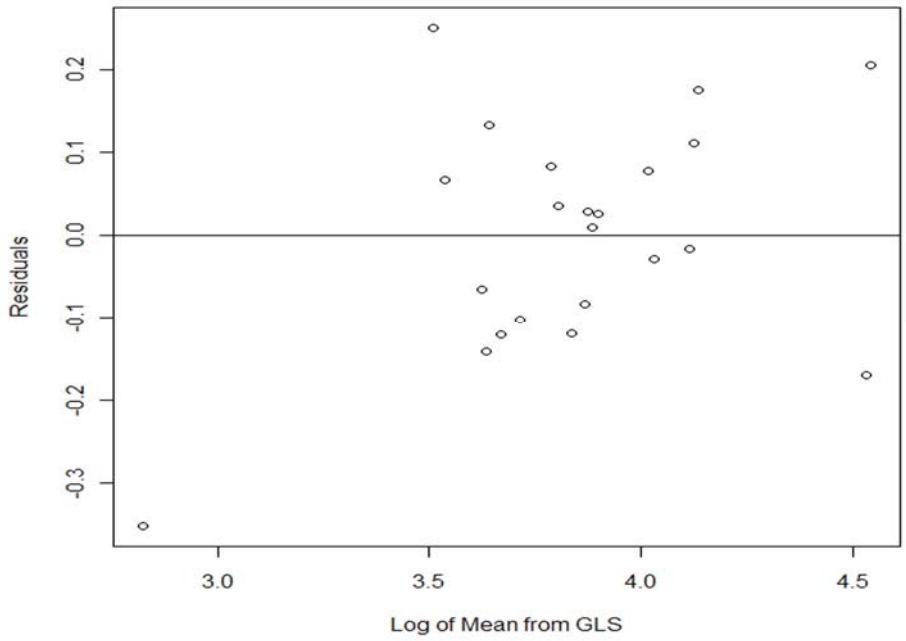


Figure D.8: Plot of residuals for GLS regression model of the mean for Region 2.

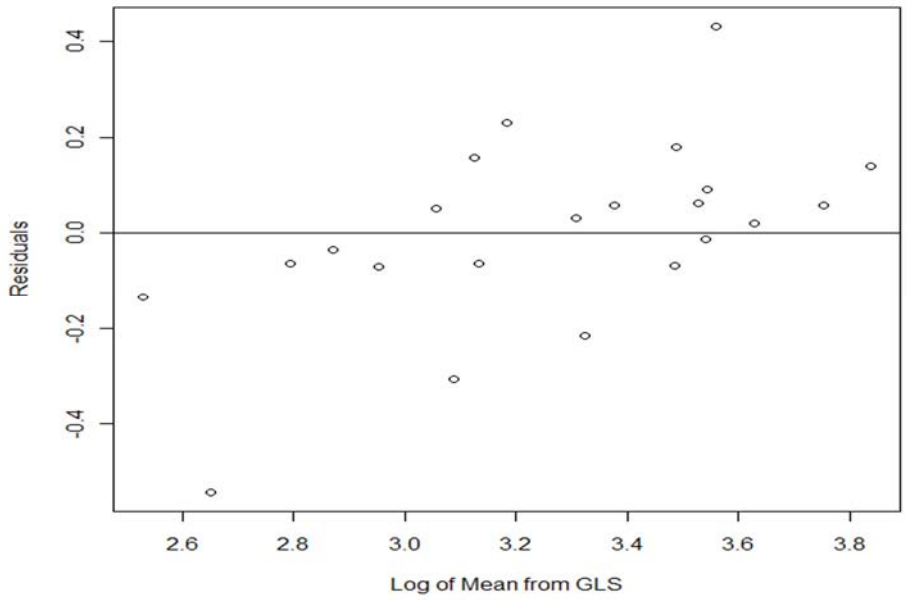


Figure D.9: Plot of residuals for GLS regression model of the mean for Region 3.

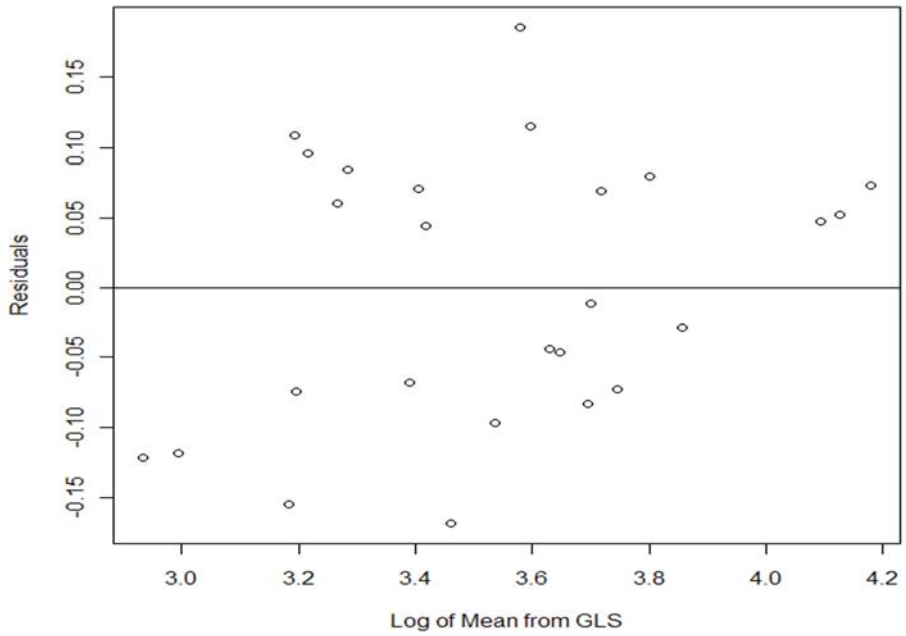


Figure D.10: Plot of residuals for GLS regression model of the mean for Region 4.

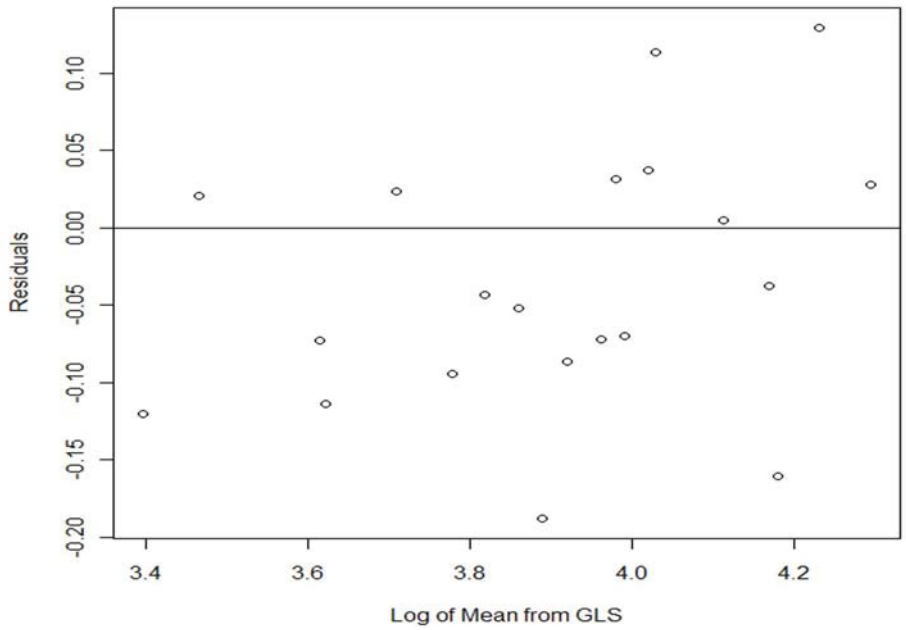


Figure D.11: Plot of residuals for GLS regression model of the mean for Region 5.

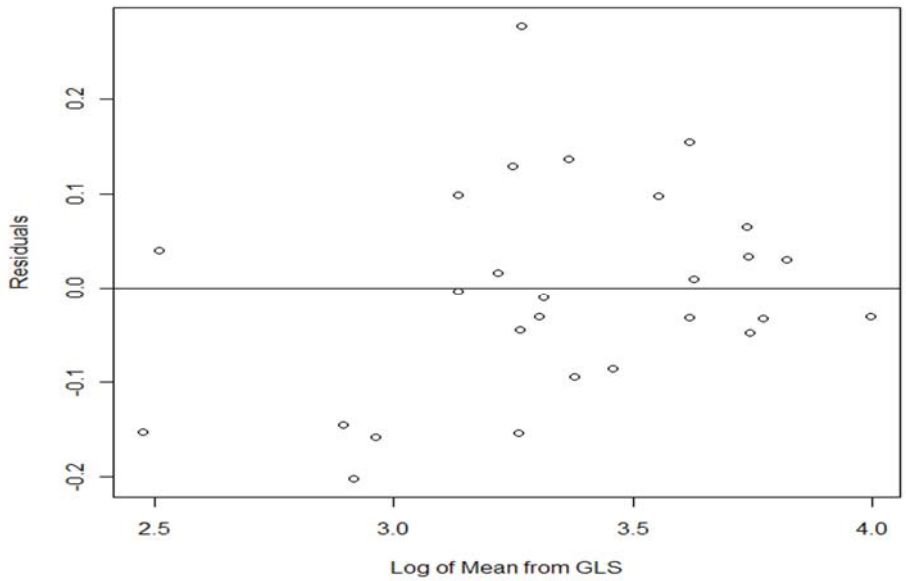


Figure D.12: Plot of residuals for GLS regression model of the mean for Region 6.

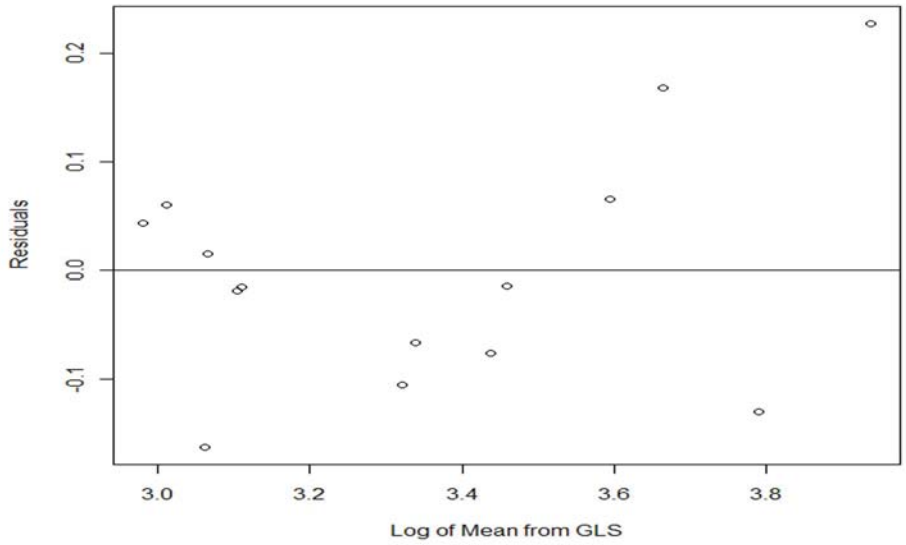


Figure D.13: Plot of residuals for GLS regression model of the mean for Region 7.

Appendix E

Additional Analyses for Classification of Ungauged Sites

This appendix presents additional information used to assess the classification scheme used to allocated ungauged sites to a region within Set 1 as proposed in Chapter 5. Herein, the characteristics of sites within each region in Set 1 are compared to the characteristics of sites from Set 2 allocated to each region in Set 1 using Wilcoxon-Mann-Whitney tests, and confusion matrices for alternative classification schemes are presented.

E1. Wilcoxon-Mann-Whitney Tests

Table E.1-Table E.7 report summary statistics for each of the nine physical variables computed for sites within each of the seven regions of Set 1, as well as for sites from Set 2 allocated to each region of Set 1 using the classification scheme based on LD_1. For each variable, a Wilcoxon-Mann-Whitney test was performed to assess the significance of the physical differences between the sites in Set 1 and those from Set 2 allocated to the same region. The tests were performed using the "wilcox.test" function in R. The resulting p-values are presented in Table E.8; the table also contains the results of tests on the precipitation as summarized for each region in Table 5.11. Overall, no significant differences are observed with respect to area at the 5% level, and very few significant differences are observed with respect to channel slope, elevation, percent forest cover, and percent impervious surface. Most of the significant differences are observed with respect to basin slope, soil drainage and precipitation, which may explain

why application of the GLS-IF approach explained in Chapter 5 does not produce quantile estimates for sites in Set 2 with as much precision as those derived for sites in Set 1. While, some of these differences are accounted for by using the proposed scheme to weight the estimated mean by precipitation, it is possible that additional gains in quantile precision could be achieved by using additional metrics to account for differences in basin slope and soil drainage, especially because these are two of the key physical variables used as indicators of extreme hydrologic response.

Table E.1
Summary statistics for the nine physical variables computed for sites within Region 1 of Set 1 versus that of sites from Set 2 allocated therein.

	A	**S_{Ch}**	**S_B**	**S_h**	**E**	**Imp**	**F**	**SI**	**Inf**
	Set 1								
Minimum	14.0	7.81	26.8	2.8	2070.2	0.1	79.1	2.3	2.0
Maximum	655.0	230.9	50.7	11.4	4056.5	5.6	99.2	3.1	3.5
Average	179.2	53.5	37.2	6.5	2875.3	1.3	88.7	2.8	2.3
Std. Dev.	186.2	60.7	7.0	2.1	632.6	1.7	7.2	0.2	0.4
	Set 2								
Minimum	15.9	4.7	14.9	1.6	2092.2	0	47.9	2.3	2.0
Maximum	812.2	189.0	46.7	30.7	3978.5	1.9	99.0	3.1	3.6
Average	164.9	50.2	33.7	9.6	3030.0	0.6	79.6	2.8	2.3
Std. Dev.	179.4	46.2	8.9	7.9	488.9	0.5	14.1	0.2	0.4

Table E.2
Summary statistics for the nine physical variables computed for sites within Region 2 of Set 1 versus that of sites from Set 2 allocated therein.

	A	**S_{Ch}**	**S_B**	**S_h**	**E**	**Imp**	**F**	**SI**	**Inf**
	Set 1								
Minimum	24.4	3.3	11.4	2.4	1174.4	0	0	2.9	2.0
Maximum	2914	149.3	37.4	26.9	1940.0	10.0	88.3	3.3	2.8
Average	288.5	41.2	20.4	7.9	1587.1	3.3	67.5	3.0	2.4
Std. Dev.	607.2	41.6	6.4	5.9	234.7	3.3	23.7	0.1	0.2
	Set 2								
Minimum	26.0	4.2	17.3	3.6	1367.8	0.2	37.8	2.8	2.1
Maximum	875.0	81.2	33.6	16.9	2496.0	5.7	90.1	3.3	3.2
Average	199.7	23.9	25.1	8.7	1627.2	1.3	69.9	3.0	2.4
Std. Dev.	228.5	19.6	5.2	3.6	281.5	1.4	13.4	0.2	0.3

Table E.3
Summary statistics for the nine physical variables computed for sites within Region 3 of Set 1 versus that of sites from Set 2 allocated therein.

	A	**S_{Ch}**	**S_B**	**S_h**	**E**	**Imp**	**F**	**SI**	**Inf**
	Set 1								
Minimum	16.8	4.5	3.7	2.7	334.9	0	0	2.6	1.9
Maximum	291.8	37.0	15.5	8.3	600.5	10.3	83.5	3.4	2.3
Average	110.5	11.0	6.2	5.0	505.5	2.9	55.2	3.1	2.1
Std. Dev.	80.1	7.7	2.5	1.4	75.2	2.6	24.6	0.2	0.1
	Set 2								
Minimum	5.0	3.6	5.8	3.9	365.3	0.3	41.5	3.0	2.1
Maximum	1372	23.5	14.2	12.1	682	3.9	68.7	3.3	2.7
Average	246.9	8.9	8.4	7.4	555.6	1.4	59.0	3.2	2.3
Std. Dev.	351.3	5.7	2.5	2.6	90.5	1.0	7.7	0.1	0.2

Table E.4
Summary statistics for the nine physical variables computed for sites within Region 4 of Set 1 versus that of sites from Set 2 allocated therein.

	A	**S_{Ch}**	**S_B**	**S_h**	**E**	**Imp**	**F**	**SI**	**Inf**
	Set 1								
Minimum	9.6	2.6	4.3	2.8	793.7	0	0	2.9	2.0
Maximum	756.0	22.6	16.0	28.3	1119.8	40.5	60.6	3.3	3.2
Average	165.3	9.2	8.2	7.1	910	9.5	38.0	3.1	2.3
Std. Dev.	178.3	5.9	3.1	5.2	92.5	11.4	21.7	0.1	0.3
	Set 2								
Minimum	9.2	5.4	6.0	3.3	705.8	0.6	8.3	3.0	2.1
Maximum	2587	35.4	15.4	13.9	1173.5	34.3	65.9	3.3	3.0
Average	281.8	11.5	9.8	6.9	905.1	6.4	44.5	3.1	2.3
Std. Dev.	607.7	6.9	3.1	3.2	132.2	9.3	16.1	0.1	0.3

Table E.5
Summary statistics for the nine physical variables computed for sites within Region 5 of Set 1 versus that of sites from Set 2 allocated therein.

	A	**S_{Ch}**	**S_B**	**S_h**	**E**	**Imp**	**F**	**SI**	**Inf**
	Set 1								
Minimum	47.1	2.0	3.6	4.0	36.3	0.2	46.7	2.9	2.1
Maximum	792.1	21.4	15.5	21.8	184.4	2.7	83.5	3.4	2.9
Average	293.5	7.1	7.6	9.0	87.6	1.0	60.5	3.1	2.4
Std. Dev.	185.5	4.6	3.5	4.8	42.8	0.7	9.8	0.2	0.2
	Set 2								
Minimum	2.7	0.7	0.3	2.5	26.6	0.1	15.7	2.8	1.9
Maximum	1236	18.0	3.4	8.1	121.8	1.9	63.4	5.9	3.9
Average	270.8	4.1	1.9	5.7	69.6	0.8	36.4	4.5	2.9
Std. Dev.	365.5	4.0	0.8	1.7	29.9	0.6	12.0	0.9	0.5

Table E.6
Summary statistics for the nine physical variables computed for sites within Region 6 of Set 1 versus that of sites from Set 2 allocated therein.

	A	S_{Ch}	S_B	S_h	E	Imp	F	SI	Inf
	Set 1								
Minimum	9.8	1.7	0.9	2.7	201.9	0	0	3.4	2.2
Maximum	1250	14.4	3.8	17.8	420.2	11.3	52.4	5.4	3.4
Average	246.5	4.9	2.3	7.1	302.1	5.7	34.0	4.0	2.6
Std. Dev.	291.6	3.2	0.8	3.7	70.3	2.2	11.0	0.5	0.3
	Set 2								
Minimum	29.8	2.2	2.5	4.2	153.3	0.2	25.4	3.0	2.1
Maximum	1228	13.2	12.6	19.1	398.2	9.7	72.9	4.3	2.9
Average	313.2	5.9	6.1	11.6	293.6	2.5	52.6	3.3	2.3
Std. Dev.	310.0	3.4	1.9	4.8	69.8	2.3	13.1	0.3	0.2

Table E.7
Summary statistics for the nine physical variables computed for sites within Region 7 of Set 1 versus that of sites from Set 2 allocated therein.

	A	S_{Ch}	S_B	S_h	E	Imp	F	SI	Inf
	Set 1								
Minimum	20.8	0.6	0	2.7	66.4	0	0	3.4	2.2
Maximum	1129	737.5	0.8	7.4	235.4	8.9	52.8	6.2	3.8
Average	323.6	255.0	0.2	4.9	141.7	2.6	37.4	5.1	3.2
Std. Dev.	348.5	239.3	0.2	1.4	51.9	3.1	15.5	1.0	0.6
	Set 2								
Minimum	49.0	1.6	2.1	5.8	129.5	1.2	21.1	3.0	2.0
Maximum	676.0	10.7	6.5	17.4	158.2	1.8	72.7	4.5	2.7
Average	302.8	5.6	4.1	9.5	139.8	1.5	41.9	3.7	2.4
Std. Dev.	287.4	4.6	2.1	5.4	13.2	0.3	23.6	0.7	0.3

Table E.8
p-values of Wilcoxon-Mann-Whitney test.
Bold indicates that the null hypothesis is rejected at 5% significance level.

	Region						
	1	**2**	**3**	**4**	**5**	**6**	**7**
A	0.887	0.790	0.181	0.852	0.091	0.493	1.00
S_{Ch}	0.984	0.258	0.355	0.134	**0.004**	0.241	0.232
S_B	0.372	**0.014**	**0.000**	**0.064**	**0.000**	**0.000**	**0.003**
S_h	0.967	0.164	**0.004**	0.664	**0.015**	**0.001**	**0.037**
E	0.220	0.965	**0.038**	0.740	0.228	0.774	0.915
Imp	0.670	0.156	0.130	0.419	0.419	**0.000**	0.595
F	0.064	0.609	0.733	0.463	**0.000**	**0.000**	1.00
SI	0.714	0.108	**0.001**	**0.002**	**0.000**	**0.000**	**0.037**
Inf	0.477	0.390	**0.000**	0.489	**0.002**	**0.000**	0.061
Pr	**0.000**	**0.000**	**0.001**	**0.000**	**0.000**	**0.003**	**0.014**

E2. Confusion Matrices for Alternative Classification Schemes

The confusion matrix for the classification scheme based on the first discriminant function as employed in Chapter 5 is presented in Table 5.8. Therein, it was observed that the classification scheme works well, except in Regions 5 and 7. The tables below present the confusion matrices for alternative classification schemes based on higher order discriminant functions. These results indicate that the success of the classification of ungauged sites could be improved by employing both the first and second discriminant functions, which together explain roughly 93% of the differences between the regions (see Table 5.6). Little or no gain in classification success would be achieved by retaining additional discriminant functions.

Table E.9
Confusion matrix for the classification scheme based on the first two discriminant functions.

	Region						
Actual\Prediction	1	2	3	4	5	6	7
Region 1	12	0	0	0	0	0	0
Region 2	1	21	0	0	0	0	0
Region 3	0	0	22	0	0	0	0
Region 4	0	0	0	26	0	0	0
Region 5	0	0	0	0	20	0	0
Region 6	0	0	0	0	0	26	1
Region 7	0	0	0	0	0	0	14
Prediction Accuracy (%):	92	100	100	100	100	100	93

Table E.10
Confusion matrix for the classification scheme based on the first three discriminant functions.

	Region						
Actual\Prediction	1	2	3	4	5	6	7
Region 1	12	0	0	0	0	0	0
Region 2	1	21	0	0	0	0	0
Region 3	0	0	22	0	0	0	0
Region 4	0	0	0	26	0	0	0
Region 5	0	0	0	0	20	0	0
Region 6	0	0	0	0	0	26	1
Region 7	0	0	0	0	0	0	14
Prediction Accuracy (%):	92	100	100	100	100	100	93

Table E.11
Confusion matrix for the classification scheme based on the first four discriminant functions.

	Region						
Actual\Prediction	1	2	3	4	5	6	7
Region 1	12	0	0	0	0	0	0
Region 2	1	21	0	0	0	0	0
Region 3	0	0	22	0	0	0	0
Region 4	0	0	1	25	0	0	0
Region 5	0	0	0	0	20	0	0
Region 6	0	0	0	0	0	27	0
Region 7	0	0	0	0	0	1	13
Prediction Accuracy (%):	92	100	96	100	100	96	100

Table E.12
Confusion matrix for the classification scheme based on the first five discriminant functions.

	Region						
Actual\Prediction	1	2	3	4	5	6	7
Region 1	12	0	0	0	0	0	0
Region 2	1	21	0	0	0	0	0
Region 3	0	0	22	0	0	0	0
Region 4	0	0	1	25	0	0	0
Region 5	0	0	0	0	20	0	0
Region 6	0	0	0	0	0	27	0
Region 7	0	0	0	0	0	1	13
Prediction Accuracy (%):	92	100	96	100	100	96	100

Table E.13

Confusion matrix for the classification scheme based on all six discriminant functions.

	Region						
Actual\Prediction	1	2	3	4	5	6	7
Region 1	12	0	0	0	0	0	0
Region 2	1	21	0	0	0	0	0
Region 3	0	0	22	0	0	0	0
Region 4	0	0	1	25	0	0	0
Region 5	0	0	0	0	20	0	0
Region 6	0	0	0	0	0	27	0
Region 7	0	0	0	0	0	1	13
Prediction Accuracy (%):	92	100	96	100	100	96	100